

Psychometrika

VOLUME XVI—1951

JANUARY-DECEMBER

Editorial Council

Chairman:—L. L. THURSTONE *Managing Editor:*—
DOROTHY C. ADKINS
Editors:—A. K. KURTZ
M. W. RICHARDSON *Assistant Managing Editor:*—
SAMUEL B. LYERLY

Editorial Board

R. L. ANDERSON	CHARLES M. HARSH	GEORGE E. NICHOLSON
J. B. CARROLL	PAUL HORST	M. W. RICHARDSON
H. S. CONRAD	ALSTON S. HOUSEHOLDER	P. J. RULON
L. J. CRONBACH	TRUMAN L. KELLEY	WM. STEPHENSON
E. E. CURETON	ALBERT K. KURTZ	GODFREY THOMSON
MAX D. ENGELHART	IRVING LORGE	R. L. THORNDIKE
HENRY E. GARRETT	QUINN MCNEMAR	L. L. THURSTONE
J. P. GULFORD	FREDERICK MOSTELLER	LEDYARD TUCKER
HAROLD GULLIKSEN		S. S. WILKS

PUBLISHED QUARTERLY

By THE PSYCHOMETRIC SOCIETY
AT 23 WEST COLORADO AVENUE
COLORADO SPRINGS, COLORADO

P

VOL
NUM
M A
1 9

26

Psychometrika

A JOURNAL DEVOTED TO THE DEVELOPMENT OF PSYCHOLOGY AS A QUANTITATIVE RATIONAL SCIENCE

THE PSYCHOMETRIC SOCIETY

• ORGANIZED IN 1935

VOLUME 16
NUMBER 1
MARCH
1951

PSYCHOMETRIKA, the official journal of the Psychometric Society, is devoted to the development of psychology as a quantitative rational science. Issued four times a year, on March 15, June 15, September 15, and December 15

MARCH 1951, VOLUME 16, NUMBER 1

Printed for the Psychometric Society at 23 West Colorado Avenue, Colorado Springs, Colorado. Entered as second class matter, September 17, 1940, at the Post Office of Colorado Springs, Colorado, under the act of March 3, 1879. Editorial Office, Department of Psychology, The University of North Carolina, Chapel Hill, North Carolina.

Subscription Price: The regular subscription rate is \$10.00 per volume. The subscriber receives each issue as it comes out, and a second complete set for binding at the end of the year. All annual subscriptions start with the March issue and cover the calendar year. All back issues are available. The price is \$1.25 per issue or \$5.00 per volume (one set only). Members of the Psychometric Society pay annual dues of \$5.00, of which \$4.50 is in payment of a subscription to *Psychometrika*. Student members of the Psychometric Society pay annual dues of \$3.00, of which \$2.70 is in payment for the journal.

Application for membership and student membership in the Psychometric Society, together with a check for dues for the calendar year in which application is made, should be sent to

RAYMOND A. KATZELL, Chairman of the Membership Committee
Psychological Services Center, Syracuse University, Syracuse 10, New York

Payments: All bills and orders are payable in advance. Checks covering membership dues should be made payable to the *Psychometric Society*. Checks covering regular subscription to *Psychometrika* and back issue orders should be made payable to the *Psychometric Corporation*. All checks, notices of change of address, and business communications should be addressed to

ROBERT L. THORNDIKE, Treasurer, Psychometric Society and Psychometric Corporation
Teachers College, Columbia University
New York 27, New York

Articles on the following subjects are published in *Psychometrika*:

- (1) the development of quantitative rationale for the solution of psychological problems;
- (2) general theoretical articles on quantitative methodology in the social and biological sciences;
- (3) new mathematical and statistical techniques for the evaluation of psychological data;
- (4) aids in the application of statistical techniques, such as nomographs, tables, work-sheet layouts, forms, and apparatus;
- (5) critiques or reviews of significant studies involving the use of quantitative techniques.

The emphasis is to be placed on articles of type (1), in so far as articles of this type are available.

In the selection of the articles to be printed in *Psychometrika*, an effort is made to obtain objectivity of choice. All manuscripts are received by one person, who

(Continued on the back inside cover page)

Psychometrika

CONTENTS

CHARLES ISAAC MOSIER, 1910-1951 - - - - -	1
HAROLD GULLIKSEN	
REMARKS ON THE METHOD OF PAIRED COMPARISONS: I. THE LEAST SQUARES SOLUTION ASSUMING EQUAL STANDARD DEVIATIONS AND EQUAL CORRELATIONS - - - - -	3
FREDERICK MOSTELLER	
THE DIMENSIONS OF TEMPERAMENT - - - - -	11
L. L. THURSTONE	
A NOTE ON CORRECTING FOR CHANCE SUCCESS IN OBJECTIVE TESTS - - - - -	21
SAMUEL B. LYERLY	
CLASSIFICATION BY MULTIVARIATE ANALYSIS - -	31
T. W. ANDERSON	
THE RELATIONSHIP BETWEEN THE METHOD OF SUCCESSIVE RESIDUALS AND THE METHOD OF EXHAUSTION - - - - -	51
WILBUR L. LAYTON	
THE RELATIONSHIP BETWEEN THE VALIDITY OF A SINGLE TEST AND ITS CONTRIBUTION TO THE PREDICTIVE EFFICIENCY OF A TEST BATTERY	57
PAUL HORST	
AN EMPIRICAL VERIFICATION OF THE WHERRY-GAYLORD ITERATIVE FACTOR ANALYSIS PROCEDURE - - - - -	67
ROBERT J. WHERRY, JOEL T. CAMPBELL, AND ROBERT PERLOFF	

(Continued)

BF
1
P98

THE CENTRAL INTELLECTIVE FACTOR - - - -	75
H. J. A. RIMOLDI	
ON THE STANDARD LENGTH OF A TEST - - - -	103
MAX A. WOODBURY	
THE ESTIMATION OF THE PARAMETERS OF A NEGA- TIVE BINOMIAL DISTRIBUTION WITH SPECIAL REFERENCE TO PSYCHOLOGICAL DATA - -	107
HERBERT S. SICHEL	
A SUCCESSIVE APPROXIMATION METHOD OF MAXI- MIZING TEST VALIDITY - - - - -	129
GOLDINE C. GLESER AND PHILIP H. DuBOIS	
JOHN VON NEUMANN AND OSKAR MORGENSTERN, <i>The- ory of Games and Economic Behavior</i> - - - -	141
A Review by Walter A. Rosenblith	
BOOKS RECEIVED - - - - -	147
PSYCHOMETRIC MONOGRAPHS ANNOUNCEMENT -	149

*General
Hapson*

PSYCHOMETRIKA—VOL. 16, NO: 1
MARCH, 1951

Charles Isaac Mosier 1910-1951

The death of Charles Mosier after a very short illness is a great loss to his friends, and a severe blow to work in the field of psychometrics. He was taken ill at his office on Monday, January 15, and died in the hospital the next afternoon with a diagnosis of meningitis.

As Chief of Research and Analysis in the Personnel Research Section, The Adjutant General's Office, he held one of the key positions for directing military research in this country; he strove to make the numerous projects under his direction contribute both to the development of valuable military tools, and to the furtherance of psychology.

He was born in Miami, Florida on June 11, 1910, and completed his college work at the University of Florida in 1932. The same year he was awarded an S. S. R. C. Fellowship and began his graduate work in psychology at the University of Chicago.

In the summer of 1933, Mary F. Fortis and he were married. Their daughter, Mary Fortis, was born a year later. He interrupted his graduate work to accept a position as Instructor in Psychology at the University of Florida in 1933. During the next four years he continued as instructor at Florida, and completed his work for the Ph.D. at Chicago.

His doctoral thesis was on a multiple-factor analysis of neurotic symptoms. This was one of the early studies applying the methods of factor analysis to items, in an attempt to discover the structure of a non-cognitive domain. From 1937 to 1941 Dr. Mosier continued his work as assistant professor and as a staff member of the Examiner's Office at the University of Florida. His papers dealt with various factor problems, such as the effects of random error, and the development of new methods of rotation. During this time he published his articles on the duality of psychophysics and test theory, which mark a new and very intriguing approach to problems in both fields.

In 1941 he went to Washington, D. C. to work for the State Technical Advisory Service, Social Security Board, first as a Research Psychologist, and later as Chief of Methods and Analysis and Chief of

Research and Test Construction. In these positions he was concerned with the development and validation of tests used for the selection of employees working for various federal and state agencies which administered the social security program.

In 1946 he moved to the Civilian Personnel Division of the Office of the Secretary of War and in 1947 to the Personnel Research Section of the AGO. Since then, as Chief, Research and Analysis, he directed a staff of psychologists and assistants in the Pentagon, and has also aided in guiding the numerous outside psychological research projects financed by the AGO. It is fortunate that a person of his ability and research acumen was available for a key position such as this. He was one of the founders of the journal *Personnel Psychology*, and was on the editorial boards of *Psychometrika* and *Educational and Psychological Measurement*. However, his attention to directive work of this sort was a loss in that it meant that he had less time to devote to his own original contributions to psychological theory—in particular, his important treatment of the duality of test theory and psychophysics.

His many friends extend their sympathy to his wife and daughter. They will remember Charlie Mosier not only as a contributor to psychology, but also as a cheerful and energetic companion and worker who carried more than his share in any enterprise.

HAROLD GULLIKSEN
Princeton University

REMARKS ON THE METHOD OF PAIRED COMPARISONS:
I. THE LEAST SQUARES SOLUTION ASSUMING
EQUAL STANDARD DEVIATIONS
AND EQUAL CORRELATIONS*

FREDERICK MOSTELLER
HARVARD UNIVERSITY

Thurstone's Case V of the method of paired comparisons assumes equal standard deviations of sensations corresponding to stimuli and zero correlations between pairs of stimuli sensations. It is shown that the assumption of zero correlations can be relaxed to an assumption of equal correlations between pairs with no change in method. Further the usual approach to the method of paired comparisons Case V is shown to lead to a least squares estimate of the stimulus positions on the sensation scale.

1. *Introduction.* The fundamental notions underlying Thurstone's method of paired comparisons (4) are these:

- (1) There is a set of stimuli which can be located on a subjective continuum (a sensation scale, usually not having a measurable physical characteristic).
- (2) Each stimulus when presented to an individual gives rise to a sensation in the individual.
- (3) The distribution of sensations from a particular stimulus for a population of individuals is normal.
- (4) Stimuli are presented in pairs to an individual, thus giving rise to a sensation for each stimulus. The individual compares these sensations and reports which is greater.
- (5) It is possible for these paired sensations to be correlated.
- (6) Our task is to space the stimuli (the sensation means), except for a linear transformation.

*This research was performed in the Laboratory of Social Relations under a grant made available to Harvard University by the RAND Corporation under the Department of the Air Force, Project RAND.

There are numerous variations of the basic materials used in the analysis—for example, we may not have n different individuals, but only one individual who makes all comparisons several times; or several individuals may make all comparisons several times; the individuals need not be people.

Furthermore, there are “cases” to be discussed—for example, shall we assume all the intercorrelations equal, or shall we assume them zero? Shall we assume the standard deviations of the sensation distributions equal or not?

The case which has been discussed most fully is known as Thurstone's Case V. Thurstone has assumed in this case that the standard deviations of the sensation distributions are equal and that the correlations between pairs of stimulus sensations are zero. We shall discuss a standard method of ordering the stimuli for this Case V. Case V has been employed quite frequently and seems to fit empirical data rather well in the sense of reproducing the original proportions of the paired comparison table. The assumption of equal standard deviations is a reasonable first approximation. We will not stick to the assumption of zero correlations, because this does not seem to be essential for Case V.

2. *Ordering Stimuli with Error-Free Data.* We assume there are a number of objects or stimuli, O_1, O_2, \dots, O_n . These stimuli give rise to sensations which lie on a single sensation continuum S . If X_i and X_j are single sensations evoked in an individual I by the i th and j th stimuli, then we assume X_i and X_j to be jointly normally distributed for the population of individuals with

$$\begin{aligned} \text{mean of } X_i &= S_i & (i = 1, 2, \dots, n) \\ \text{variance of } X_i &= \sigma^2(X_i) = \sigma^2 & (i = 1, 2, \dots, n) \\ \text{correlation of } X_i \text{ and } X_j &= \rho_{ij} = \rho & (i, j = 1, 2, \dots, n). \end{aligned} \quad (1)$$

The marginal distributions of the X_i 's appear as in Figure 1.

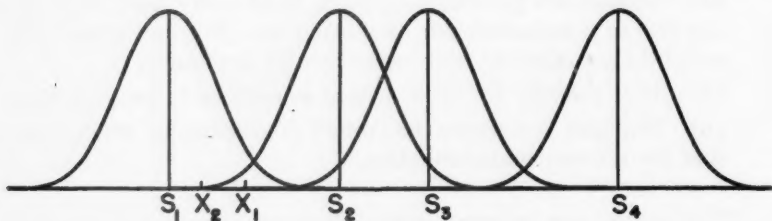


FIGURE 1

The Marginal Distributions of the Sensations Produced by the Separate Stimuli in Thurstone's Case V of the Method of Paired Comparisons.

The figure indicates the possibility that $X_2 < X_1$, even though $S_1 < S_2$. In fact this has to happen part of the time if we are to build anything more than a rank-order scale.

An individual I compares O_i and O_j and reports whether $X_i \geq X_j$ (no ties are allowed).

We can best see the tenor of the method for ordering the stimuli if we first work through the problem in the case of nonfallible data. For the case of nonfallible data we assume we know the true proportion of the time X_i exceeds X_j , and that the conditions given above (1) are exactly fulfilled.

Our problem is to find the spacing of the stimuli (or the spacing of the mean sensations produced by them, the $S_1 \dots S_n$ points in Figure 1). Clearly we cannot hope to do this except within a linear transformation, for the data reported are merely the percentages of times X_i exceeds X_j , say p_{ij} .

$$p_{ij} = P(X_i > X_j) = \frac{1}{\sqrt{2\pi}\sigma(d_{ij})} \int_0^\infty e^{-\frac{[d_{ij} - (S_i - S_j)]^2}{2\sigma^2(d_{ij})}} dd_{ij} \quad (2)$$

where $d_{ij} = X_i - X_j$, and $\sigma^2(d_{ij}) = 2\sigma^2(1 - \rho)$. There will be no loss in generality in assigning the scale factor so that

$$2\sigma^2(1 - \rho) = 1. \quad (3)$$

It is at this point that we depart slightly from Thurstone, who characterized Case V as having equal variances and zero correlations. However, his derivations only assume the correlations are zero explicitly (and artificially), but are carried through implicitly with equal correlations (not necessarily zero). Actually this is a great easing of conditions. We can readily imagine a set of attitudinal items on the same continuum correlated .34, .38, .42, i.e., nearly equal. But it is difficult to imagine them all correlated zero with one another. Past uses of this method have all benefited from the fact that items were not *really* assumed to be uncorrelated. It was only *stated* that the model assumed the items were uncorrelated, but the model was unable to take cognizance of the statement. Guttman (2) has noticed this independently.

With the scale factor chosen in equation (3), we can rewrite equation (2)

$$p_{ij} = \frac{1}{\sqrt{2\pi}} \int_{-(S_i - S_j)}^{\infty} e^{-1/2 y^2} dy. \quad (4)$$

From (4), given any p_{ij} we can solve for $-(S_i - S_j)$ by use of a normal table of areas. Then if we arbitrarily assign as a location parameter $S_1 = 0$, we can compute all other S_i . Thus given the p_{ij} matrix we can find the S_i . The problem with fallible data is more complicated.

3. *Paired Comparison Scaling with Fallible Data.* When we have fallible data, we have p'_{ij} which are estimates of the true p_{ij} . Analogous to equation (4) we have

$$p'_{ij} = \frac{1}{\sqrt{2\pi}} \int_{-D'_{ij}}^{\infty} e^{-1/2 y^2} dy, \quad (5)$$

where the D'_{ij} are estimates of $D_{ij} = S_i - S_j$. We merely look up the normal deviate corresponding to p'_{ij} to get the matrix of D'_{ij} . We notice further that the D'_{ij} need not be consistent in the sense that the D_{ij} were; i.e.,

$$D_{ij} + D_{jk} = S_i - S_j + S_j - S_k = D_{ik},$$

does not hold for the D'_{ij} .

We conceive the problem as follows: from the D'_{ij} to construct a set of estimates of the S_i 's called S'_i , such that

$$\sum_{i,j} [D'_{ij} - (S'_i - S'_j)]^2 \text{ is to be a minimum.} \quad (6)$$

It will help to indicate another form of solution for nonfallible data. One can set up the $S_i - S_j$ matrix:

MATRIX OF $S_i - S_j$

	1	2	3	n
1	$S_1 - S_1$	$S_1 - S_2$	$S_1 - S_3$		$S_1 - S_n$
2	$S_2 - S_1$	$S_2 - S_2$	$S_2 - S_3$		$S_2 - S_n$
3	$S_3 - S_1$	$S_3 - S_2$	$S_3 - S_3$		$S_3 - S_n$
.					
.					
.					
n	$S_n - S_1$	$S_n - S_2$	$S_n - S_3$		$S_n - S_n$
Totals	$\sum S_i - nS_1$	$\sum S_i - nS_2$	$\sum S_i - nS_3$		$\sum S_i - nS_n$
Means	$\bar{S} - S_1$	$\bar{S} - S_2$	$\bar{S} - S_3$		$\bar{S} - S_n$

Now by setting $S_1 = 0$, we get $S_2 = (\bar{S} - S_1) - (\bar{S} - S_2)$, $S_3 = (\bar{S} - S_1) - (\bar{S} - S_3)$, and so on. We will use this plan shortly for the S'_i .

If we wish to minimize expression (6) we take the partial derivative with respect to S'_i . Since $D'_{ij} = -D'_{ji}$ and $S'_i - S'_j = -(S'_j - S'_i)$ and $D'_{ii} = S'_i - S'_i = 0$, we need only concern ourselves with the sum of squares from above the main diagonal in the $D'_{ij} - (S'_i - S'_j)$ matrix, i.e., terms for which $i < j$. Differentiating with respect to S'_i we get:

$$\frac{\partial(\Sigma/2)}{\partial S'_i} = 2 \left[\sum_{j=1}^{i-1} (D'_{ji} - S'_j + S'_i) - \sum_{\substack{j=i+1 \\ (i=1, 2, \dots, n)}}^n (D'_{ij} - S'_i + S'_j) \right] \quad (7)$$

Setting this partial derivative equal to zero we have

$$\begin{aligned} &+S'_1 + S'_2 \dots + S'_{i-1} - (n-1)S'_i + S'_{i+1} + \dots + S'_n \\ &= \sum_{j=1}^{i-1} D'_{ji} - \sum_{j=i+1}^n D'_{ij} \quad (i=1, 2, \dots, n), \end{aligned} \quad (8)$$

but $D'_{ij} = -D'_{ji}$, and $D'_{ii} = 0$; this makes the right side of (8)

$$\sum_{j=1}^{i-1} D'_{ji} + D'_{ii} + \sum_{j=i+1}^n D'_{ji} = \sum_{j=1}^n D'_{ji}.$$

Thus (8) can be written

$$\sum_{j=1}^n S'_j - nS'_i = \sum_{j=1}^n D'_{ji} \quad (i=1, 2, \dots, n). \quad (9)$$

The determinant of the coefficients of the left side of (9) vanishes. This is to be expected because we have only chosen our scale and have not assigned a location parameter. There are various ways to assign this location parameter, for example, by setting $\bar{S}' = 0$ or by setting $S'_1 = 0$. We choose to set $S'_1 = 0$. This means we will measure distances from S'_1 . Then we try the solution (10) which is suggested by the similarity of the left side of (9) to the total column in the matrix of $S_i - S_j$.

$$S'_i = \sum_{j=1}^n D'_{j1}/n - \sum_{j=1}^n D'_{ji}/n. \quad (10)$$

Notice that when $i=1$, $S'_i = 0$ and that

$$\sum_{i=1}^n S'_i = \sum_{i=1}^n D'_{j1}$$

because

$$\sum_i \sum_j D'_{ji} = 0,$$

which happens because every term and its negative appear in this double sum. Therefore, substituting (10) in the left side of (9) we have

$$\sum_{i=1}^n D'_{j1} - n \left[\sum_{j=1}^n D'_{j1}/n - \sum_{j=1}^n D'_{ji}/n \right] = \sum_{j=1}^n D'_{ji}, \quad (11)$$

which is an identity, and the equations are solved. Of course, any linear transformation of the solutions is equally satisfactory.

The point of this presentation is to provide a background for the theory of paired comparisons, to indicate that the assumption of zero correlations is unnecessary, and to show that the customary solution to paired comparisons is a least squares solution in the sense of condition (6). That this is a least squares solution seems not to be mentioned in the literature although it may have been known to Horst (3), since he worked closely along these lines.

This least squares solution is not entirely satisfactory because the p'_{ij} tend to zero and unity when extreme stimuli are compared. This introduces unsatisfactorily large numbers in the D'_{ij} table. This difficulty is usually met by excluding all numbers beyond, say, 2.0 from the table. After a preliminary arrangement of columns so that the S'_i will be in approximately proper order, the quantity

$$\sum (D'_{ij} - D'_{i,j+1})/k$$

is computed where the summation is over the k values of i for which entries appear in both column j and $j+1$. Then differences between such means are taken as the scale separations (see for example Guilford's discussion (1) of the method of paired comparisons). This method seems to give reasonable results. The computations for methods which take account of the differing variabilities of the p'_{ij} and therefore of the D'_{ij} seem to be unmercifully extensive.

It should also be remarked that this solution is not entirely a reasonable one because we really want to check our results against the original p'_{ij} . In other words, a more reasonable solution might

be one such that once the S'_i are computed we can estimate the p'_{ij} by p''_{ij} , and minimize, say,

$$\sum (p'_{ij} - p''_{ij})^2$$

or perhaps

$$\sum (\arcsin \sqrt{p'_{ij}} - \arcsin \sqrt{p''_{ij}})^2.$$

Such a thing can no doubt be done, but the results of the author's attempts do not seem to differ enough from the results of the present method to be worth pursuing.

REFERENCES

1. Guilford, J. P. *Psychometric Methods*. New York: McGraw-Hill Book Co., 1936, 227-8.
2. Guttman, L. An approach for quantifying paired comparisons and rank order. *Annals of math. Stat.*, 1946, 17, 144-163.
3. Horst, P. A method for determining the absolute affective values of a series of stimulus situations. *J. educ. Psychol.*, 1932, 23, 418-440.
4. Thurstone, L. L. Psychophysical analysis. *Amer. J. Psychol.*, 1927, 38, 368-389.

Manuscript received 8/22/50

1871

...

...

...

...

...

THE DIMENSIONS OF TEMPERAMENT*

L. L. THURSTONE
THE UNIVERSITY OF CHICAGO

The correlations among the thirteen personality scores yielded by the Guilford schedule for factors STDCR, and the Guilford-Martin schedules for factors GAMIN, and O, Ag, and Co, as reported by Lovell, were factored by the centroid method. The purpose was to see how many factors were represented by the thirteen scores; therefore the test reliabilities were used in the diagonal cells. It was found that the scores represent not more than nine linearly independent factors. The orthogonal factor matrix was rotated to oblique simple structure. Seven of the oblique factors were given tentative interpretation. Two factors were regarded as residual factors because of the small variance which they represent. The seven factors have been named Active, Vigorous, Impulsive, Dominant, Stable, Sociable, and Reflective.

The purpose of this study was to determine the number of factors or dimensions that are implied in current personality schedules, and also to ascertain the nature of each factor or type. The several schedules of Guilford were chosen for this purpose because they represent careful analytical work. Each of his schedules has previously been analyzed factorially, and correlations have been determined between the separate scores for his schedules.

The various personality schedules cover a wide range of personal characteristics, including those which are relatively permanent for each person as well as those which change more or less from one year to the next because of social experience. Most of the scores derived from the Guilford schedules represent relatively permanent characteristics of a person which may be called temperamental traits. Some personality scores, such as appraisals of attitudes on controversial social questions, represent only partly the temperamental characteristics of a person. Such scores also reflect his recent social experience, his social identifications, and the propaganda to which he may have been exposed. They are less stable as indicators of temperamental types. Our interest here is in those non-intellective traits of personality which are relatively stable, the temperamental types, and which are

*This study was supported in part by a research grant from Sears Roebuck and Company. The writer wishes to acknowledge in particular the interest and assistance of Mr. J. C. Worthy of the National Personnel Department at Sears Roebuck and Company. The writer also wishes to acknowledge the assistance of Mr. James Degan who was responsible for the computing in this study.

not often markedly changed in social experience. Hence we refer to this problem as the dimensions of temperament rather than the much larger domain that is called personality.

Guilford has produced three personality schedules that were used in the present study. These were Guilford's schedule for the scores STDCR, the Guilford-Martin schedule for the scores GAMIN, and the Guilford-Martin schedule for the scores O, Ag, and Co.* Each of the first two schedules gives five scores, and the third schedule gives three scores. Hence the schedules give thirteen separate scores, all of which were used in the present study.

The correlations among the thirteen scores were reported recently by Lovell who gave all three schedules to 213 subjects.† She made a factor analysis of the thirteen scores in which the communalities were determined by their intercorrelations. This is the usual procedure, but in the present case it should be recalled that the thirteen scores were themselves determined as factor scores from the original questionnaires that contained many hundreds of items. Hence the procedure of Lovell was essentially to investigate the second-order domain in the thirteen factor scores. This is an interesting and important problem. The second-order domain in the traits of temperament may be psychologically revealing. But before undertaking such a study, it would be preferable to make sure that the factor scores which enter into a second-order analysis are linearly independent. Lovell questions the linear independence of the thirteen scores in her opening statement. She says: "The original studies showed that the thirteen factors were not completely independent of each other though they were sufficiently separate to make individual scores helpful." Test scores may be very useful even though they are not linearly independent, but such a situation introduces reservations about a second-order analysis.

In the present study we direct ourselves first to the main problem, namely, to determine the number of dimensions or factors in these personality schedules which are represented by thirteen separate scores. This is the same problem that Lovell mentions in introducing her study. Instead of dealing with the thirteen scores as variables whose common factors are to be ascertained, we want to know how many factors are represented in the thirteen scores. For this purpose we make the factorial analysis with the test reliabilities in the diagonal.

*Guilford, J. P., and Guilford, R. B. Personality factors, S, E, and M, and their measurement. *J. Psychol.*, 1936, 2, 107-127; Personality factors, D, R, T, and A. *J. abnorm. soc. Psychol.*, 1939, 34, 21-36; Personality factors N and GD. *J. abnorm. soc. Psychol.*, 1939, 34, 239-248.

†Lovell, Constance. A study of factor structure of thirteen personality variables. *Educ. psychol. Meas.*, 1945, 5, 335-350.

al cells. If a second-order analysis is to be made of these thirteen scores, then the common factor variances, the communalities, are recorded in the diagonal cells as was done in Lovell's paper.

The thirteen scores from the Guilford Schedules are listed in Table 1. Each trait is shown by Guilford's name for the trait and by his code symbol. Then follow some items indicative of the presence of the trait (positive items) and some items that indicate the absence of the trait (negative items). Then follow further sample items from the schedules. Most of Guilford's scores are defined by a mixture of positive and negative items. A few of the scores have a preponderance of negative items in which a subject gets a high score in a trait by acknowledging the opposite trait. For bipolar traits, this is legitimate, especially when both directions are well represented by questions. When only one of the two poles is well defined by questions, it seems preferable to let the well defined pole carry the trait name even if it is not regarded as the socially preferable side of the bipolarity. We have given here Guilford's trait names and the direction of each bipolarity to which he assigns the numerically higher scores. When he plots a profile of percentile norms, he reverses some of the scores so as to represent the socially more desirable end of the bipolarity with the numerically higher percentile ranks.

In some of the schedules, there is a good balance between "yes" and "no" answers that indicate presence of the trait, but in several of the schedules there is a large majority of one type of answer for the high scores. The Cycloid score is determined from 68 items that are positively scored with "yes" answers while there are only five items with "no" answers for the same trait. The corresponding ratios for other schedules are General Activity 21 and 3, Nervousness 5 and 38, Objectivity 2 and 45, Agreeableness 2 and 36, Cooperativeness 4 and 56. Some of these traits seem to be more easily described by the socially less desirable side of the bipolarity.

The correlations reported by Lovell are reproduced in our Table 2. In the diagonal cells of this correlation matrix we have recorded the test reliabilities which are also reported by Lovell.* The question is

*The writer agrees with a reservation that has been made by one of the reviewers of this paper, but it probably does not invalidate an approximate determination of the dimensionality of the thirteen scores. The reservation is as follows:

The author might well mention two conditions that have important bearings on his analysis. The reliabilities reported by Lovell were taken from the test manuals and were therefore not based upon the same population as the intercorrelations. Accuracy of these values would be very important in establishing the dimensionality of the factors. Many of the intercorrelations are spuriously

now to determine the rank of this correlation matrix. The matrix was factored by one of the centroid methods† and the result is shown in our Table 3. In making this factorial reduction, we did not adjust the diagonal values because our object was to find the number of dimensions in the test scores and this is not necessarily the rank of the reduced correlation matrix. This objective excludes the error variance in each test score but we do not limit ourselves to the factors that may be common to the test scores. We want to analyze the dimensionality of the test content of the thirteen scores, excluding their error variance.

The orthogonal centroid factor matrix of Table 3 shows nine factors. The distribution of ninth factor residuals is shown in Table 4. We now have the answer to our first problem in the result that, for practical purposes, the thirteen personality scores represent not more than nine factors. Hence the thirteen scores are linearly dependent.

A good structure was obtained in solving the rotational problems for these data. The transformation matrix Λ is shown in Table 5, by which the nine orthogonal centroid axes are replaced by seven oblique axes which have been given interpretation and by two residual axes. The last two axes are given only tentative interpretation and they may be left as residual factors because of the small variance which they represent.

In Table 6 we have the oblique factor matrix for the seven factors with interpretation and two residual factors that are denoted X_1 and X_2 respectively. The thirteen scores from the Guilford schedules are represented here in the socially favorable forms in accordance with the scoring on Guilford's profiles and as they are represented in Lovell's correlation matrix. We shall now give tentative interpretation to the seven significant primary factors of this V matrix.

In the first column there are eight zero loadings and five significant loadings. The strongest of these is Thinking Introversion with .76. The next strongest saturation is $-.41$ on Rhathymia.‡ The three less conspicuous loadings are Social Introversion with .29, Depression with .35, and Emotional Instability with .26. Inspection of the items leads to the generalization that this primary factor can be called *introversion* or *introspection*. A more general descriptive adjective is

high due to the fact that some items were scored for more than one trait. Such scores therefore have common error variance which adds materially to the apparent common-factor variance.

†Thurstone, L. L. Multiple factor analysis. Chicago: Univ. of Chicago Press, 1947. Chap. VIII, pp. 161-170.

‡The signs have been adjusted to agree with the reversal of trait names.

Reflective which covers most of the traits in a descriptive sense without implying socially favorable or unfavorable implications. The primary factor is denoted *R*.

The second column has only two large significant saturations, namely, those for *Agreeableness* and *Cooperativeness* with small saturations for *Objectivity*, *Freedom from Nervousness*, and *Rhathymia*. We have generalized this primary factor in the descriptive adjective *Sociable* with the symbol *S*.

The third column has two significant saturations, namely, those for *Emotional Stability* and *Freedom from Depression* with lower saturations on *Social Extraversion*, *Thinking Extraversion*, and on *Freedom from Nervousness*. We generalize this primary factor in the descriptive adjective *Emotionally Stable*, with the symbol *E*.

The next column has only one large saturation, namely, that for *Masculinity* with a smaller saturation on *Freedom from Nervousness*. This primary factor can be generalized in the adjective *Vigorous* with the symbol *V*.

The next column has two significant saturations on *Ascendancy* and *Extraversion* with no other significant loadings. We have generalized this primary factor in the descriptive adjective *Dominant* in the sense of social leadership with the symbol *D*.

The next column has two significant saturations on *General Activity* and *Cooperativeness* with slight saturations on *Objectivity* and *Ascendancy*. We have generalized this primary factor in the descriptive adjective *Active* with the symbol *A*.

The seventh column has only two significant entries, namely, those for *General Activity* and for *Rhathymia*. A small saturation on *Freedom from Inferiority Feelings* is consistent with the generalization of this primary factor in the descriptive adjective *Impulsive* with the symbol *I*.

The residual factor X_1 might be called *Self-confidence* and we were tempted to denote it *C*, but the saturations are small with the highest loading of .35; and it seemed more appropriate not to include it in a list of primary traits until its independence and significance can be demonstrated more clearly. The second residual factor X_2 was also left without interpretation since its highest saturation is only .29 on *Freedom from Nervousness*.

In Table 7 we have the intercorrelations of the seven primary factors to which we have attempted to give interpretation. The most conspicuous intercorrelation is that of *Impulsiveness* and *Dominance* which is .71. These two factors are clearly separated in the oblique factor matrix *V* so that by the present data they are clearly distin-

guished. Another significant correlation between primaries is that of *Sociability* and *Emotional Stability* which is .52. The separation between these two primaries is also clear in the oblique factor matrix *V*.

The thirteen scores obtained from the several personality schedules of Guilford represent a dimensionality of not more than nine linearly independent factors. Since the variance of two of these factors is rather small, the dimensionality of the thirteen scores is not more than seven independent factors for practical purposes. The analysis was made in terms of the test space from which only the error variance was eliminated. Hence the reliability coefficients were used in the diagonal cells of the correlation matrix for this analysis. A similar analysis with communalities in the diagonals would probably give a smaller number of dimensions since it would be limited to those factors which are shared by two or more of the thirteen scores. That was not the purpose of the present study.

The seven dimensions of the thirteen scores for which interpretation has been attempted were tentatively named Reflective (introspective), Impulsive, Sociable, Active, Dominant (leadership), Vigorous, and Emotionally Stable. These primary factors were given the symbols *R*, *I*, *S*, *A*, *D*, *V*, and *E*, respectively. The simple structure that was found in this configuration of seven dimensions was very marked, as shown in the large number of vanishing entries in the oblique factorial matrix *V*. The structure can be seen even more clearly on a diagram for each pair of columns in which the saturations of one column are plotted against the corresponding saturations in the other column.

This variant interpretation of Guilford's work on personality factors does not deny the existence of many more factors in this domain. The thirteen traits that are described and named by Guilford can be very useful even though they are not linearly independent. In general, there would be preference for a set of descriptive profile categories in which each column contributes some information that cannot be obtained as a weighted score of the other columns. That is, of course, what is meant by linear independence.

We started this analysis with the expectation of finding bipolar factors for all or most of these factors; but the result revealed all of them to be positive. In naming the factors we tried to avoid those terms which refer explicitly to the more abnormal aberrations of temperament or personality, such as depression and cycloid disposition. Such concepts refer to the psychiatric extremes, but they have correlates in terms that refer to the less severe deviations within the normal range of temperament. When schedules of this kind are used for the description of personality among subjects who are in the normal range, it

seems preferable to use terms which avoid as far as possible the comparison of a normal subject with the abnormal extremes. This is probably good policy in describing the temperaments of normal subjects even though it is recognized that there is no sharp demarcation between the normal and the abnormal in each of the factors or dimensions.

Manuscript received 5/16/50

Revised manuscript received 10/9/50

TABLE 1
Guilford's Thirteen Scores

S: Social Introversion

Positive: Shyness, seclusiveness, tendency to withdraw from social contacts.

Negative: Sociability, tendency to seek social contacts, to enjoy company of others.

Sample items: Limits acquaintances to a select few, keeps quiet in social groups, difficulty in starting conversation with strangers, frequent loneliness, spends evenings alone, takes life seriously, bashfulness, lets others take the lead.

T: Thinking Introversion

Positive: Inclination to meditative or reflective thinking, philosophizing, analyzing one's self and others.

Negative: Extravertive orientation in thinking.

Sample items: Analyzes motives of others, ponders over the past, takes life seriously, works on complicated problems, often lost in thought, much attention to details, often moody, works better when praised.

D: Depression

Positive: Habitually gloomy, pessimistic mood, feelings of guilt.

Negative: Cheerfulness and optimism.

Sample items: Often moody, self-conscious, daydreams frequently, often worries, frequent ups and downs in mood, feelings easily hurt, loneliness, difficulty in making decisions, feelings of inferiority, often excited.

C: Cycloid

Positive: Strong emotional fluctuations, tendency toward flightiness, emotional instability.

Negative: Uniformity in mood, evenness of disposition.

Sample items: Moody, acts on the spur of the moment, works better when praised, changes work frequently, daydreams, worries, ups and downs in mood, feelings easily hurt, impulsive, interests change quickly, lonely, high-strung, absent-minded.

R: Rathymia

Positive: Happy-go-lucky, carefree disposition, lively, impulsive.

Negative: Inhibited, over-controlled, conscientious, serious-minded.

TABLE 1 (Continued)
Guilford's Thirteen Scores

Sample items: Carefree, acts on spur of the moment, impulsive, craves excitement, jumps at conclusions, lively, plays pranks on others, restless.

G: General Activity

Positive: General pressure for vigorous activity.

Sample items: Quick in actions, eats rapidly, walks fast, "on the go," starts work with enthusiasm, hurries, talkative, impulsive, daredevil, group leader.

A: Ascendancy

Positive: Social Leadership

Sample items: Easily starts conversation with strangers, good at bluffing, organizer, takes social initiative, likes public speaking, takes responsibility, takes charge in case of accident, stands up for his rights, a good salesman.

M: Masculinity

Positive: Masculinity in emotional and temperamental make-up.

Sample items: Wants to be physically strong, not afraid of the dark, likes hunting, likes to take a chance, not afraid of deep water, not sorry for underdog, not afraid of snakes, preference for mathematics, science, politics, building trades, mining, prize fights, rather than literature, music, flowers, art, dancing.

I: Inferiority Feelings

Positive: Lack of confidence, undervaluation of one's self, feelings of inadequacy.

Sample items: Often feels thwarted, bossed around too much, often bored, slow emotional recovery from emotional upset, awkward, craves encouragement, absent-minded, unpopular, easily discouraged, slow in making decisions.

N: Nervousness

Positive: Jumpiness, jitteriness, easily distracted, irritated, easily annoyed.

Negative: Calm, unruffled, relaxed.

O: Lack of objectivity

Positive: Takes everything personally, hypersensitive, easily upset, nervous, disturbed by criticism, readily unburdens his troubles to others, easily offended, or annoyed.

Ag: Lack of Agreeableness

Positive: Does not like to take instructions from others, feels that most people are stupid, hates to lose an argument, dislikes many people, takes pleasure in bossing people, selfish, frequently in conflict, contempt for opinions of others, self-confident about his own abilities, "hard-boiled."

Co: Lack of cooperativeness

Positive: Lack of faith in people, believes most people shirk their duties, dislikes his superiors, against large business corporations, dislikes traffic regulations, distrustful of all successful people.

TABLE 2
Correlation Matrix for Guilford's Thirteen Scores

		S	T	D	C	R	G	A	M	I	N	O	Ag	Co
1.	S	.90	.42	.64	.44	.66	.38	.73	.10	.59	.38	.47	.14	.22
2.	T	.42	.84	.65	.59	.30	-.07	.20	.21	.34	.39	.41	.17	.24
3.	D	.64	.65	.94	.90	.23	-.04	.48	.32	.74	.71	.75	.34	.44
4.	C	.44	.59	.90	.88	-.02	-.19	.31	.33	.68	.70	.72	.35	.42
5.	R	.66	.30	.23	-.02	.90	.56	.53	.04	.27	.08	.21	-.08	-.02
6.	G	.38	-.07	-.04	-.19	.56	.89	.44	-.07	.09	-.23	-.06	-.31	-.17
7.	A	.73	.20	.48	.31	.53	.44	.88	.26	.57	.33	.46	.00	.20
8.	M	.10	.21	.32	.33	.04	-.07	.26	.85	.33	.35	.37	.01	.21
9.	I	.59	.34	.74	.68	.27	.09	.57	.33	.91	.67	.75	.35	.45
10.	N	.38	.39	.71	.70	.08	-.23	.33	.35	.67	.89	.72	.47	.53
11.	O	.47	.41	.75	.72	.21	-.06	.46	.37	.75	.72	.83	.50	.62
12.	Ag	.14	.17	.34	.35	-.08	-.31	.00	.01	.35	.47	.50	.80	.63
13.	Co	.22	.24	.44	.42	-.02	-.17	.20	.21	.45	.53	.62	.63	.91

TABLE 3
Orthogonal Factor Matrix *F*

	I	II	III	IV	V	VI	VII	VIII	IX
1.	.75	-.47	.18	.05	.20	-.13	-.07	-.11	.09
2.	.58	.13	.58	.10	-.26	-.04	.03	.11	-.15
3.	.88	.19	.28	-.14	.10	-.02	.07	-.06	.08
4.	.77	.41	.28	-.25	.13	.04	.15	-.02	.04
5.	.45	-.70	.13	.25	-.16	.16	-.22	.14	.05
6.	.15	-.79	-.17	-.12	-.18	.25	.27	-.14	.04
7.	.67	-.50	-.17	-.17	.14	-.29	.06	.16	-.08
8.	.41	.20	-.22	-.38	-.50	-.22	-.32	.09	.07
9.	.83	.07	-.15	-.18	.19	.14	-.08	.11	.14
10.	.74	.39	-.08	-.09	.13	.12	-.16	-.09	-.18
11.	.83	.25	-.14	.03	.08	.09	.06	.17	.07
12.	.42	.44	-.22	.51	.23	.08	-.05	-.11	-.08
13.	.58	.38	-.35	.42	-.08	-.14	.21	-.11	.13

TABLE 4
Frequency Distribution of 9th Factor Residuals

	<i>N</i> = 156								
Residual:	-.05	-.04	-.03	-.02	-.01	.00	.01	.02	.03
Frequency:	2	2	8	22	26	42	32	12	10

TABLE 5
Transformation Matrix A

	R	S	E	V	D	A	I	X ₁	X ₂
I	.26	.30	.30	.17	.16	.16	.16	.15	.12
II	-.06	.13	.10	.08	-.28	-.09	-.32	.03	.04
III	.80	-.40	.48	-.21	-.03	-.34	-.15	-.17	-.03
IV	.34	.80	-.28	-.23	.01	.08	.04	-.28	.08
V	-.42	-.20	-.04	-.56	.37	-.34	-.32	.28	-.15
VI	-.02	.08	.07	-.11	-.85	-.01	.85	.18	.14
VII	.03	-.02	.00	-.72	-.04	.84	.11	.31	-.06
VIII	.06	-.22	-.76	-.17	-.04	-.11	-.01	.82	-.17
IX	.00	.00	.00	.00	-.16	.00	.00	.00	-.95

TABLE 6
Oblique Factor Matrix V

	Guilford's Scores	R	S	E	V	D	A	I	X ₁	X ₂
1-S	Social Extraversion	-.29	.11	.32	.01	.42	-.01	.06	-.03	-.04
2-T	Thinking Extraversion	-.76	.06	.36	.07	.00	-.01	.02	-.02	.22
3-D	Freedom from Depression	-.35	.05	.50	.04	.12	.05	-.01	.13	.01
4-C	Emotional Stability	-.26	-.05	.50	-.02	.00	.05	-.05	.22	.02
5-R	Rhathymia	-.41	.21	-.03	.14	.07	-.04	.45	-.02	.03
6-G	General Activity	.02	.00	.05	-.07	-.04	.44	.60	.02	.01
7-A	Ascendancy	.03	-.02	-.03	.03	.55	.18	.00	.30	.04
8-M	Masculinity	.00	.00	.08	.74	.01	-.01	-.04	.01	.02
9-I	Freedom from Inferiority									
	Feelings	.05	.12	.15	.14	.04	.02	.17	.35	-.06
10-N	Freedom from Nervousness	-.01	.24	.32	.24	.00	-.07	.05	.10	.29
11-O	Objectivity	-.08	.31	.07	.06	.00	.16	.13	.34	.02
12-Ag	Agreeableness	.03	.66	.00	-.05	-.01	.03	-.03	-.06	.19
13-Co	Cooperativeness	-.03	.72	.00	.03	.07	.43	-.03	-.03	.00

TABLE 7
Correlations between Primary Factors

	R	S	E	V	D	A	I	X ₁	X ₂
R	1.00	-.11	-.23	.15	.07	.11	-.01	.06	-.02
S	-.11	1.00	.52	-.03	.01	-.37	-.15	.56	-.14
E	-.23	.52	1.00	.05	.04	-.18	-.10	.66	-.12
V	.15	-.03	.05	1.00	.03	.32	-.11	.30	-.09
D	.07	.01	.04	.03	1.00	-.17	.71	.03	-.19
A	.11	-.37	-.18	.32	-.17	1.00	-.26	-.16	.04
I	-.01	-.15	-.10	-.11	.71	-.26	1.00	-.19	-.22
X ₁	.06	.56	.66	.30	.03	-.16	-.19	1.00	-.01
X ₂	-.02	-.14	-.12	-.09	-.19	.04	-.22	-.01	1.00

A NOTE ON CORRECTING FOR CHANCE SUCCESS IN OBJECTIVE TESTS

SAMUEL B. LYERLY

THE UNIVERSITY OF NORTH CAROLINA

The conventional scoring formula to "correct for guessing" is derived and is compared with a regression method for scoring which has been recently proposed by Hamilton. It is shown that the usual formula, $S = R - W/(n-1)$, yields a close approximation (correct within one point) to the maximum-likelihood estimate of an individual's "true score" on the test, if we assume that the individual "knows" or "does not know" the answer to each item, that guessing at unknown items is random, and that success at guessing is governed by the binomial law. It is also shown that the usual scoring formula yields an unbiased estimate of the individual's "true score," when the true score is defined as the mean score over an indefinitely large number of independent attempts at the test or at equivalent (parallel) tests.

Hamilton's recent paper on correcting for guessing in objective tests* proposes a method which differs greatly from that which most test technicians have customarily used. Hamilton advocates the use of a regression equation based upon the known or assumed distribution of examinee knowledge, i.e., the distribution of scores which would be obtained if guessing were excluded and each individual answered only those items which he "knew" and refrained from marking those which he did not know. Assuming a binomial distribution of examinee knowledge, and assuming further that every examinee makes a response to every item and that the relative frequency of successful responses on those items whose answers are not known will be governed by the binomial law, Hamilton derives an equation of the form

$$S_i = (k\bar{R} - n)R_i / (k-1)\bar{R}, \quad (1)$$

where

*Hamilton, C. Horace. Bias and error in multiple-choice tests. *Psychometrika*, 1950, 15, 151-168.

- S_i = the estimated true score for Individual i ,
 k = the number of alternatives per item,
 R_i = the raw score (number of items correctly answered) for Individual i ,
 \bar{R} = the mean raw score for the group of N individuals, and
 n = the number of items in the test.

The factor $(k\bar{R} - n)/(k - 1)\bar{R}$ in Eq. (1) is the regression coefficient of true scores on raw scores for the group of individuals under consideration. Hamilton shows that this regression is in general non-linear, depending upon the form of the distribution of examinee knowledge, and departing from linearity as the distribution of true scores departs from the binomial. He presents alternative formulas for use when some or all of the examinees do not complete the test and when the assumed distribution of examinee knowledge is not binomial.

Hamilton considers that the conventional scoring formula, which is of the form

$$S_i = (kR_i - n)/(k - 1), \quad (2)$$

is a mistaken one, based upon the regression of raw scores upon true scores.

It is the purpose of this note to present a complete derivation of the conventional scoring formula and to show that under the assumptions of (1) perfect knowledge or perfect ignorance on each item of the test, (2) "pure" random guessing at unknown items, and (3) a binomial distribution of success at guessing on unknown items (assumptions which Hamilton also makes), the usual scoring method yields a value which is a close approximation to the maximum-likelihood estimate of the desired true score and is, in addition, an unbiased estimate of the true score defined as the limit approached by the mean estimated score in an indefinitely large number of attempts at the same test (or parallel tests).

Test technicians have not generally considered the problem of correcting for guessing as a regression problem in the ordinary sense of the term, but as a problem in *sampling*. We have a sample of responses—the test responses of Individual i —and we wish to estimate therefrom a "true score;" i.e., we wish to determine the true score which would make the obtained raw score most probable. We are

thus making an inference from sample to population. It must be noted that the "population" here is the class of true scores to which Individual i 's true score belongs—not the class of true scores of all members of some group of individuals. Similarly, the "sample" under consideration is the single set of test responses of Individual i rather than a set of scores made by a sample of individuals. These distinctions are fundamental for a comparison of Hamilton's scoring system with the conventional one.

The test user who employs the correction formula may have either of two purposes in mind: (1) He may wish to estimate the exact number of answers the subject knows *on a single test*, or (2) he may wish to estimate the *mean* "true score" of the subject in a series of trials at the test (or at parallel tests). We shall consider the two problems in order.

1. *Estimating the Exact Number of Answers Known on a Single Test*

Let us suppose that Individual i makes a raw score R_i on a test of n items each having k alternatives. We shall assume that R_i is the sum of an unknown number S_i , which represents the number of items to which he knows the answers, and an unknown number ($R_i - S_i$) of items for which he receives credit as a consequence of lucky random guesses. We wish to estimate S_i under the assumption that success at guessing on the $(n - S_i)$ unknown items is binomially distributed. (Since we are dealing with one individual, we shall omit the subscript.)

There are $R + 1$ hypotheses which we may entertain concerning the value of S . We may list them as follows:

$H(S = R)$: The subject knew the answers to R items, guessed at the remaining $(n - R)$ items, and was unsuccessful in each case. $S = R$.

$H(S = R - 1)$: The subject knew the answers to $(R - 1)$ items and guessed at the remaining $(n - R + 1)$. He was successful in one guess and unsuccessful in $(n - R)$. $S = R - 1$.

.....
 $H(S = 0)$: The subject did not know any of the answers and guessed at all n items. He was successful in R guesses, unsuccessful in $(n - R)$. $S = 0$.

For each of the hypotheses we can compute a probability based upon the binomial law:

$$\left. \begin{aligned} P(S=R) &= [(n-R)!/0!(n-R)!]p^0q^{n-R} \\ P(S=R-1) &= [(n-R+1)!/1!(n-R)!]p^1q^{n-R} \\ &\vdots \\ P(S=0) &= [n!/R!(n-R)!]p^Rq^{n-R} \end{aligned} \right] \quad (3)$$

The hypothesis with the greatest probability as calculated by Eqs. (3) is the one which we will accept, in the sense that we consider it the one with the greatest likelihood of being the "true" hypothesis. As an illustration, let us consider the case of 5 correct answers out of 10 in a 5-response multiple-choice test. There are 6 hypotheses which may be examined: $S=5$, $S=4$, $S=3$, $S=2$, $S=1$, and $S=0$. Applying Eqs. (3), we have:

$$P(S=5) = \frac{5!}{0!5!} (.2)^0 (.8)^5 = .328.$$

$$P(S=4) = \frac{6!}{1!5!} (.2)^1 (.8)^5 = .393.$$

$$P(S=3) = \frac{7!}{2!5!} (.2)^2 (.8)^5 = .275.$$

$$P(S=2) = \frac{8!}{3!5!} (.2)^3 (.8)^5 = .147.$$

$$P(S=1) = \frac{9!}{4!5!} (.2)^4 (.8)^5 = .066.$$

$$P(S=0) = \frac{10!}{5!5!} (.2)^5 (.8)^5 = .026.$$

The probabilities listed above are interpreted as follows: If the hypothesis $S=5$ were true, we would expect an R of 5 with a relative frequency of about 33%; i.e., our subject through random guessing at the 5 items he did not know would be unsuccessful in all 5 questions in about one-third of an extended series of independent attempts at the test. Similarly, an obtained R of 5 which includes one successful guess in 6 attempts would be expected to occur with a relative frequency of 39% in an extended series of guessing on 6 unknown items. Since this is the largest of the computed values, we

accept the hypothesis $S=4$ as that which makes the obtained R the "most probable." (It should be noted that in making this estimate we are considering only *integral values* of S . Under our initial assumptions, we cannot entertain hypotheses of fractional S 's.)

By making a simple approximation we can arrive at an easier estimate. (In a later paragraph we shall consider the error involved in our approximation.) Each equation in (3) is one term of a binomial expansion, and the largest computed value by Eqs. (3) is at the mode of the binomial frequency distribution of which it is a term, provided that the distribution has a unique mode and provided R is sufficiently large relative to n that we do not need to consider negative S 's ($R > np$). If we accept the mean of the binomial, $S + (n - S)p$, as an estimate of its mode,* we conclude that the value of S which maximizes the probability of an obtained R is that value of S for which

$$R = S + (n - S)p.$$

Then

$$S = (R - np) / (1 - p) = (kR - n) / (k - 1). \quad (4)$$

Eq. (4) is identical with (2) above, and is the familiar formula to "correct for guessing." It is a "maximum-likelihood" estimate when S is an integer in that no other hypothesis concerning the value of S can give a higher probability of obtaining the given R .

The scoring formula does not assume that the subject answers every item on the test. We may replace n in Eq. (2) by m , letting m stand for the number of items attempted, or we may use the familiar variation

$$S = R - W / (n - 1), \quad (5)$$

where W is the number of wrong answers. The usefulness of Eq. (5) or an equivalent is obvious, since no scoring formula to adjust for chance success will alter the relative standings of individuals in a group when everyone attempts every item. Correcting for guessing is therefore rarely used unless some or all examinees do not complete the test.

A few remarks are in order concerning the use of the conventional scoring formula to provide estimates of this kind. It will be

*The mode of the binomial, if it exists, is always within one unit of the mean—in fact, it lies between $Np - q$ and $Np + p$. If there is no unique mode, $Np - q$ and $Np + p$ are consecutive integers which represent the 2 greatest frequencies in the distribution.

recalled that in the derivation we used the mean of the binomial distribution as an estimate of its mode. Since the distributions with which we are dealing are discrete and in general non-symmetrical, Eq. (2) is not exact. A study of Eqs. (3) will reveal the fact that the probability for the hypotheses $S = (R - np)/(1 - p)$ and that for the hypothesis $S = [(R - np)/(1 - p)] + 1$ are equal when S as computed by Eq. (2) is an integer and $n > R \geq np$. This follows from the theorem in probability theory that if P_1 and P_2 are the probabilities of the most probable number of successes (actually "failures" in our problem) in m and $m + 1$ trials, respectively, then $P_1 \geq P_2$, the equality sign holding when $(m + 1)p$ is an integer. For example, in a true-false test of 10 items, the probability that an individual with a true score of 6 will receive a raw score of 8 is equal to

$$\frac{4!}{2!2!} (.5)^4 = .3750.$$

The probability that a true score of 7 would give rise to a raw score of 8 is

$$\frac{3!}{1!2!} (.5)^3 = .3750.$$

Similarly, in a 5-response multiple-choice test of 10 items, a raw score of 2 can arise with equal likelihood from a true score of either 0 or 1, since

$$\frac{10!}{2!8!} (.2)^2 (.8)^8 = \frac{9!}{1!8!} (.2)^1 (.8)^8 = .30199.$$

In such cases there is no modal value for S which will maximize the probability of R except when $R = n$, in which case the mode is at $S = n$; or $R < np$, when the mode is at $S = 0$. (This latter case is justification of the common practice of assigning an adjusted score of zero to a raw score when the corrected value by Eq. (2) falls below zero, since by Eqs. (3) $S = 0$ is the most probable true score.) Since there is no modal value when the S computed by Eq. (2) is an integer and $n > R \geq np$, we may use the formula as it stands, increase each integral S by one unit, or toss a coin in a given case to determine whether to add 1 or use the calculated S .

When S as calculated by Eq. (2) is not an integer, a similar problem arises. A consideration of Eqs. (3) along with the approxi-

mation Eq. (2) reveals that the latter introduces a bias in such cases. This bias arises from the fact that the distribution of S , i.e., the distribution of probabilities calculated by the $R + 1$ equations in (3), is discrete and has a negative skew when $p < \frac{1}{2}$. The modal value of S is in such cases always the next integer above the S calculated by Eq. (2) when the latter is a fraction. For example, in the case of 7 correct answers out of 10 in a 5-response multiple-choice test, the S calculated by the correction formula is $7 - \frac{2}{5} = 6.25$. The probabilities computed by Eqs. (3) are .4096 for $S = 6$ and .5120 for $S = 7$. Thus, although 6.25 is nearer 6 than 7, the most probable value for the corresponding true score is 7. Adjusted scores calculated by Eq. (2) which are fractions should therefore be raised to the next whole number, unless $p > \frac{1}{2}$, in which case S should be *reduced* to the next integer. This latter situation would arise if, say, a test were composed of items each having 4 alternatives of which 3 were correct, and the examinee is instructed to mark only one alternative for each item. The distribution of S would in this case be positively skewed, and the use of Eq. (2) would overestimate S .

2. *Estimating the Mean "True Score" in a Series of Independent Trials at the Test (or at Parallel Tests)*

It might appear at first that this problem, estimating an individual's mean true score over an indefinitely large number of trials, can be solved by the method of Hamilton—i.e., by using the regression of true scores upon raw scores in a sample of individuals, basing calculations upon the known or assumed distribution of true scores (provided we could somehow arrive at a satisfactory idea of the true score distribution). However, if the problem of correcting for guessing is considered as a problem in regression, the appropriate estimate of an individual's true score is the mean of a number of estimates *for that individual* rather than the mean of the estimates for a group of individuals. It is the *individual's* mean toward which his scores regress, not the group mean. An analogy may clarify the point: Suppose that we ask a person to take out of his pockets all the coins which he happens to have and to throw them onto a table. Then we ask him to report the number of coins which fell in such a way that the heads are uppermost. If we are interested in estimating the total number of coins on the table from the number showing heads, the information we have is sufficient. We would not seek to improve our estimate by asking *other* individuals to empty *their*

pockets and report the number of heads showing on their coins unless we knew that everyone had the same number of coins to start with—which would make the experiment equivalent to a repetition of the procedure with the same person.

Taking this view of the estimation problem, we return to Hamilton's derivation on pages 153 and 154 and re-define N to be the number of independent trials at the test (or at parallel tests) by Individual i rather than the number of subjects in the group. We see that Hamilton's Eq. (3),

$$S = (k\bar{R} - n) / (k - 1), \quad [\text{Hamilton's Eq. (3)}]$$

which he derives as the mean true score for a group of N individuals, turns out to be the expression for the mean estimated score of a single individual over a series of N trials. But ordinarily we have only one trial from which to make our estimate and only one R_i . This single R_i , however, is an unbiased estimate of \bar{R}_i and is the best estimate we have of \bar{R}_i in the absence of other scores *for that individual*. The substitution of R_i for \bar{R}_i in our modification of Hamilton's Eq. (3) yields the familiar scoring formula, Eq. (2). Similar substitutions in Hamilton's own recommended correction formulas, Eq. (15), page 156, and Eq. (22), page 157, also yield the conventional correction formula. The traditional scoring formula, then, yields an unbiased estimate of the individual's "true score," defined as the limit approached by the mean estimated score as the number of independent trials increases. This estimate is not adjusted to an integral value as is the single-trial estimate discussed above.

Hamilton's analysis of the problem differs fundamentally from that presented above. He apparently is seeking to minimize errors of estimate over a *sample of individuals* by estimating the true score for an examinee as the mean true score for all individuals in the sample who have the same raw score. A consequence of Hamilton's method is that a subject's estimated score depends upon the distribution of scores in the group with which he is examined and will vary as the general level of performance of the group.* Hence it

*Some idea of this dependence can be gained through applying Hamilton's basic scoring formula, Eq. (1) above, to the case of an individual who has 75 correct responses on a true-false test of 100 items. By the conventional correction formula, his adjusted score is $75 - 25 = 50$, regardless of the mean performance of the group or regardless of whether he was even examined with a group.

will ordinarily be impossible for anyone to earn a perfect score (even the person who constructed the test!) if Hamilton marks the papers, unless *everyone* makes a perfect score—in which case the test ceases to be a test in the usual meaning of the term. The conventional scoring method, in which the individual's set of responses is the sample and his own true score the desired population value, does not encounter such a difficulty. The independent variables include only the subject's performance and the structure of the test. The distribution of ability in the group (or even the existence of a group) is irrelevant.

It is true, of course, that scores generally will tend to regress toward the group mean upon retest. This phenomenon is a consequence of the less-than-perfect reliability of the testing instrument and is observed in connection with tests of all kinds, whether of the "objective" type or the "free-answer" type in which the probability of chance success is presumably zero. Regression methods (which require the standard error of measurement or the reliability coefficient) may be used to adjust scores for this effect; but the problem of reliability and the question of what adjustment, if any, should be made to test scores on account of their less-than-perfect reliability have been ignored in this paper. Reliability is considered to be a separate (though not unrelated) matter and one which requires other assumptions and other information for its treatment than does the simpler problem of adjusting for chance success.

In conclusion, it should be recognized that the assumptions employed in the correction for guessing (perfect knowledge or perfect ignorance on each item and "pure" random guessing at unknown items) are only approximately appropriate for most objective tests. "Partial" knowledge, positional response tendencies, the ability to eliminate one or more alternatives, the lack of independence among items or among alternatives within an item (as, for example, when alternatives are ordered along a continuum), varying degrees of

If we substitute several values for a group mean in Hamilton's equation, we get estimates such as:

Assumed Group Mean	Estimated Score for Individual with $R_i = 75$
90	67
80	56
70	43
60	25
50	0

"willingness to gamble"—these and other circumstances weaken the validity of our mathematical model, as the results of almost any item analysis will show. The effects of these factors cannot be determined from an inspection of raw test scores, and it is doubtful that any general scoring formulas can be found which will take them into account. It may be possible, however, to devise empirical scoring methods which are approximately valid for certain kinds of tests and for certain classes of individuals.

Manuscript received 7/14/50

Revised manuscript received 9/11/50

CLASSIFICATION BY MULTIVARIATE ANALYSIS*

T. W. ANDERSON
COLUMBIA UNIVERSITY

The problem considered is the use of a set of measurements on an individual to decide from which of several populations he has been drawn. It is assumed that in each population there is a probability distribution of the measurements. Principles for choosing the rule of classification are based on costs of misclassification. Optimum procedures are derived in general terms. If the measurements are normally distributed, the procedures use one discriminant function in the case of two populations and several discriminant functions in the cases of more populations. The numerical example given involves three normal populations.

1. *The Problem of Classification.*

The problem of classification which we shall consider arises when an investigator makes a number of measurements on an individual and wishes to classify this individual into one of several categories on the basis of these measurements. The investigator cannot identify the individual with a category directly but must make his inference from these measurements. In many cases it can be assumed that there are a finite number of categories or populations from which the individual may have come, and each population is characterized by a probability distribution of the measurements. An individual is considered as a random observation from this population. The question is: Given an individual with certain measurements, from which population did he arise?

The problem of classification may be considered as a problem of "statistical decision functions." We have a number of hypotheses, each hypothesis is that the distribution of the observation is a given one. We must accept one of these hypotheses and reject the others. If only two populations are admitted, we have an elementary problem of testing one hypothesis of a specified distribution against another.†

*Sponsored in part by the Office of Naval Research.

†The general theory described in this paper can be deduced as a special case of A. Wald's theory (9). M. A. Girshick presented some of this theory to the meeting of the Institute of Mathematical Statistics at Berkeley, June 16, 1949, in "Bayes, Minimax and Other Approaches to Multiple Classification Problems," and G. W. Brown (2) described some of the results before the American Statistical Association at Cleveland, December 27, 1948.

In some instances, the categories are specified beforehand in the sense that the probability distributions of the measurements are completely known. In other cases, the form of each distribution may be known, but the parameters of the distribution must be estimated from a sample from that population. For instance, the two populations may be multivariate normal with means, variances, and correlations unknown.

We can give an example of our problem from the field of education. Prospective students applying for admission into college are given a battery of tests. The scores on these tests for a given student form a set of measurements, x_1, \dots, x_p . The prospective student may be a member of one population consisting of those students who will successfully complete college training,* or he may be a member of the other population, those who will not complete the college course successfully. The problem is to classify each student applying for admission on the basis of his scores on the entrance examinations.

2. *Standards of Good Classification.*

To begin with we shall suppose that an individual with certain measurements (x_1, \dots, x_p) has been drawn from one of two populations, π_1 and π_2 . The properties of these two populations are specified by the probability density functions (or frequency functions), $p_1(x_1, \dots, x_p)$ and $p_2(x_1, \dots, x_p)$, respectively.† We wish to define a procedure of classifying this individual as coming from π_1 or π_2 . The set of measurements x_1, \dots, x_p can be presented as a point in a p -dimensional space. We shall divide this space into two regions, R_1 and R_2 . If the point corresponding to an individual falls in R_1 we shall say the individual was drawn from π_1 , and if it falls in R_2 we shall say he came from π_2 .

We wish to select these two regions so that we minimize on the average the bad effects of misclassification. In following a given classification procedure the statistician can make two kinds of errors. If the individual is actually from π_1 the statistician can classify him as coming from population π_2 , or if he is from π_2 the statistician may classify him as from π_1 . We need to know the relative undesirability of these two kinds of misclassification. Let the "cost" of the first type of misclassification be $C(2|1)$, and let the cost of misclassifying an individual from π_2 as from π_1 be $C(1|2)$. These costs may be

*To avoid raising the question of prediction it might be better to say that this population consists of those students with potentialities for successfully completing college training.

†Each infinite population is an idealization of the population of all possible observations.

measured in any kind of units. As we shall see later, it is only the ratio of the two costs that is important. While the statistician may not know these costs in each case, he will usually have at least a rough idea of them. The following two by two table indicates the costs of correct and incorrect classification.

		Population Drawn From	
Statistician's Decision	π_1	π_2	
	π_1	0	$C(1 2)$
	π_2	$C(2 1)$	0

We shall consider two ways of defining "minimum costs." Which definition should be used will depend on the knowledge one has of the situation. Let R denote the rule of classification; the rule implies a division of the p -dimensional observation space into the two regions R_1 and R_2 . If the observation is drawn from π_1 , the probability of correct classification, $P(1|1, R)$, is the probability of falling into R_1 , and the probability of misclassification, $P(2|1, R)$, is the probability of falling into R_2 . For instance,

$$P(1|1, R) = \int_{R_1} p_1(x_1, \dots, x_p) dx_1 \dots dx_p. \quad (1)$$

Similarly, if the observation is drawn from π_2 , the probability of correct classification is $P(2|2, R)$, the integral of $p_2(x_1, \dots, x_p)$ over R_2 , and the probability of misclassification is $P(1|2, R)$. If the observation is drawn from π_1 , there is a loss when the observation is incorrectly classified as coming from π_2 ; the expected loss, or risk, is the product of the cost of a mistake times the probability of making it,

$$r(1, R) = C(2|1)P(2|1, R). \quad (2)$$

In the same way we see that when the observation is from π_2 , the expected loss due to misclassification is

$$r(2, R) = C(1|2)P(1|2, R). \quad (3)$$

In many cases we have *a priori* probabilities of drawing an observation from one or the other population. Suppose that the probability of drawing from π_1 is q_1 and from π_2 is q_2 . Then the expected loss due to misclassification is the sum of the products of the probability of drawing from each population times the expected loss for that population. It is

$$q_1 r(1, R) + q_2 r(2, R) = q_1 C(2|1)P(2|1, R) + q_2 C(1|2)P(1|2, R). \quad (4)$$

We wish to choose our regions, R_1 and R_2 , to minimize this expected loss.

In the example mentioned earlier one "cost of misclassification" is a measure of the undesirability of starting a student through college when he will not be able to finish, and the other is a measure of the undesirability of refusing to admit a student who can complete his course. If a group of students are given the tests before the classification procedure is inaugurated and all are allowed to enter, estimates of the probabilities of drawing from the two populations are obtained from the relative frequencies of students who complete and who do not complete college.

If we do not have *a priori* probabilities of drawing from π_1 and π_2 , we cannot write down (4). Then we can only speak of the expected loss *if* the observation is drawn from π_1 or *if* the observation is drawn from π_2 . For a given procedure R , the less desirable case is to have a drawing from the population with the greater risk. A conservative principle to follow is to choose our procedure so as to minimize the maximum risk. This is the so-called *minimax principle*.

3. Procedures of Classification into One of Two Populations with Known Probability Distributions.

Now let us see how we find the regions which give us the minimum expected loss. First we shall treat the case when *a priori* probabilities of drawing from π_1 and π_2 are known. Since we have *a priori* probabilities we can define joint probabilities of drawing from a given population and observing a set of variables within given ranges. The probability that an observation come from π_1 and that the i th variate be between x_i and $x_i + dx_i$ is approximately $q_1 p_1(x_1, \dots, x_p) dx_1 \dots dx_p$. Similarly, the probability of drawing from π_2 and obtaining an observation with the i th variate falling between x_i and $x_i + dx_i$ is approximately $q_2 p_2(x_1, \dots, x_p) dx_1 \dots dx_p$. If we now have an observation x_1, \dots, x_p , the conditional probability that it comes from π_1 is

$$\frac{q_1 p_1(x_1, \dots, x_p)}{q_1 p_1(x_1, \dots, x_p) + q_2 p_2(x_1, \dots, x_p)}, \quad (5)$$

and the conditional probability that it comes from π_2 is

$$\frac{q_2 p_2(x_1, \dots, x_p)}{q_1 p_1(x_1, \dots, x_p) + q_2 p_2(x_1, \dots, x_p)}. \quad (6)$$

Suppose for a moment that $C(1|2) = C(2|1) = 1$; then the expected loss is simply the probability of misclassification. For a given observation x_1, \dots, x_p we minimize the probability of misclassification if we assign it to the population with a greater *a posteriori* probability; that is, if (5) is greater than (6) we say the observation came from π_1 , and if (5) is less than (6) we say it came from π_2 . Since the denominator of (5) is the same as that of (6) our rule for classification gives us a division of the space into R_1 and R_2 according to the following:

$$\begin{aligned} R_1: q_1 p_1(x_1, \dots, x_p) &> q_2 p_2(x_1, \dots, x_p), \\ R_2: q_1 p_1(x_1, \dots, x_p) &< q_2 p_2(x_1, \dots, x_p). \end{aligned} \quad (7)$$

R_1 consists of points satisfying the first inequality, and R_2 consists of points satisfying the second.*

If $C(2|1) \neq C(1|2)$,

$$\frac{C(2|1) q_1 p_1(x_1, \dots, x_p)}{q_1 p_1(x_1, \dots, x_p) + q_2 p_2(x_1, \dots, x_p)}$$

is the conditional expected loss if we classify an observation x_1, \dots, x_p into π_2 , and

$$\frac{C(1|2) q_2 p_2(x_1, \dots, x_p)}{q_1 p_1(x_1, \dots, x_p) + q_2 p_2(x_1, \dots, x_p)}$$

is the conditional expected loss if we classify this observation into π_1 . We minimize the expected loss for this particular observation if we classify it to obtain the lower expected loss. The regions are characterized as follows:

$$\begin{aligned} R_1: C(2|1) q_1 p_1(x_1, \dots, x_p) &> C(1|2) q_2 p_2(x_1, \dots, x_p), \\ R_2: C(2|1) q_1 p_1(x_1, \dots, x_p) &< C(1|2) q_2 p_2(x_1, \dots, x_p). \end{aligned} \quad (8)$$

We could also write this as†

*The case of equality in (7) can be neglected since we assume the density functions are such that the probability of equality is zero.

†These results were first obtained in this way by Welch (10) for the case of equal costs of misclassification.

$$\begin{aligned}
 R_1: \frac{p_1(x_1, \dots, x_p)}{p_2(x_1, \dots, x_p)} &> \frac{C(1|2)q_2}{C(2|1)q_1}, \\
 R_2: \frac{p_1(x_1, \dots, x_p)}{p_2(x_1, \dots, x_p)} &< \frac{C(1|2)q_2}{C(2|1)q_1}.
 \end{aligned}
 \tag{9}$$

This is the "Bayes solution."

These inequalities seem intuitively reasonable. If the probability of drawing from π_1 is decreased or if the cost of misclassifying into π_2 is decreased, the inequality in (9) for R_1 is satisfied by fewer points. Since the regions depend on q_1 and q_2 , the expected loss does also. The curve A in Figure 1 indicates how the expected loss may vary with q_1 (and $q_2 = 1 - q_1$).

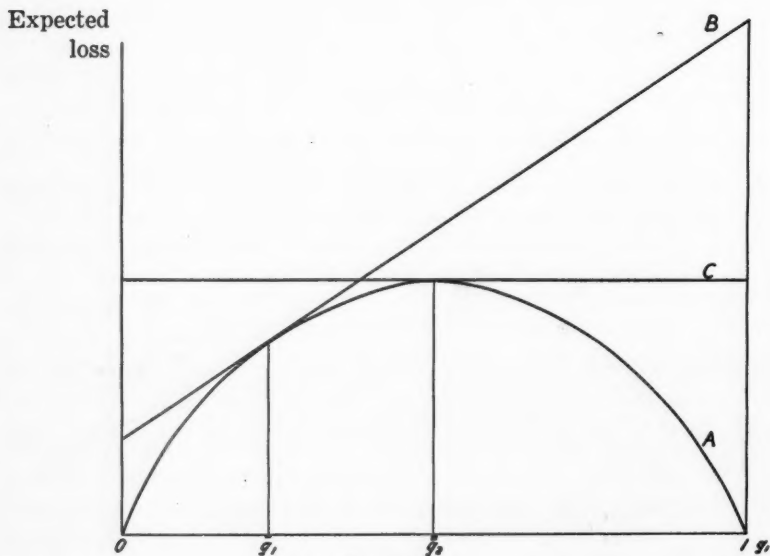


FIGURE 1

It may very well happen that the statistician errs in assigning his *a priori* probabilities. Suppose that the statistician used \bar{q}_1 and $\bar{q}_2 (= 1 - \bar{q}_1)$ when q_1^* and $q_2^* (= 1 - q_1^*)$ are the actual probabilities of drawing from π_1 and π_2 , respectively. Then the actual expected loss is

$$q_1^* C(2|1)P(2|1, \bar{R}) + (1 - q_1^*) C(1|2)P(1|2, \bar{R}), \quad (10)$$

where \bar{R}_1 and \bar{R}_2 are based on \bar{q}_1 and \bar{q}_2 . Given the regions \bar{R}_1 and \bar{R}_2 , this is a linear function of q_1^* graphed as B . This line touches A at $q_1^* = \bar{q}_1$. It cannot go below A because for $q^* \neq \bar{q}_1$ the best regions are defined by (9) for $q_1 = q_1^*$ and $q_2 = 1 - q_1^*$. From the graph we can guess that a small error in q_1 is not very important.

Now let us turn to the case where the statistician cannot assign *a priori* probabilities to the two populations. Then he may choose a procedure that minimizes the maximum expected loss. It can be shown (by the Neyman-Pearson Fundamental Lemma) that the best regions of classification are of the form

$$\begin{aligned} R_1: \frac{p_1(x_1, \dots, x_p)}{p_2(x_1, \dots, x_p)} &> k, \\ R_2: \frac{p_1(x_1, \dots, x_p)}{p_2(x_1, \dots, x_p)} &< k. \end{aligned} \quad (11)$$

where k is suitably chosen. It should be noticed that for any particular k there are *a priori* probabilities q_1 and q_2 satisfying

$$k = \frac{C(1|2)q_2}{C(2|1)q_1}. \quad (12)$$

Thus every minimax solution is a Bayes solution for some *a priori* probabilities. Since R_1 increases as k increases, and hence $r(1, R)$ increases as k increases, and at the same time $r(2, R)$ decreases, the choice of k giving the minimax solution is the one for which*

$$r(1, R) = r(2, R). \quad (13)$$

This is then the average loss for it is immaterial which population is drawn from. The graph of the risk against *a priori* probability q_1 is, therefore, a horizontal line (labelled C). Since there is one value of q_1 , say \bar{q}_1 , satisfying (12), the line C must touch A .

4. Classification into One of Two Known Multivariate Normal Populations.

One of the most interesting examples of the general theory is

*As for the Bayes solution we assume that the densities are such that the probability of equality in (11) is zero.

in the case that the populations have multivariate normal distributions with the same set of variances and correlations, but with different sets of means. Suppose that x_1, \dots, x_p have a joint normal distribution with means in π_1 of $\mathcal{E}x_i = \mu_i^{(1)}$, and in π_2 of $\mathcal{E}x_i = \mu_i^{(2)}$. Let the common set of variances and correlations be $\sigma_1^2, \dots, \sigma_p^2, \rho_{12}, \rho_{13}, \dots, \rho_{p-1,p}$. We shall find it convenient to write (11) as

$$\begin{aligned} R_1: \log \frac{p_1(x_1, \dots, x_p)}{p_2(x_1, \dots, x_p)} &> \log k, \\ R_2: \log \frac{p_1(x_1, \dots, x_p)}{p_2(x_1, \dots, x_p)} &< \log k. \end{aligned} \quad (14)$$

It is easily verified that in this particular case

$$\log \frac{p_1(x_1, \dots, x_p)}{p_2(x_1, \dots, x_p)} = \sum_{i=1}^p \lambda_i x_i - \sum_{i=1}^p \lambda_i \frac{1}{2} (\mu_i^{(1)} + \mu_i^{(2)}), \quad (15)$$

where $\lambda_1, \dots, \lambda_p$ is the solution of

$$\sum_{j=1}^p \sigma_i \sigma_j \rho_{ij} \lambda_j = \mu_i^{(1)} - \mu_i^{(2)}. \quad (16)$$

The first term in (15) is the well-known *discriminant function* obtained by Fisher (3) by choosing the linear function for which the difference in expected values for the two populations relative to the standard deviation is a maximum. The regions are given by

$$\begin{aligned} R_1: \sum_{i=1}^p \lambda_i x_i &> \sum_{i=1}^p \lambda_i \frac{1}{2} (\mu_i^{(1)} + \mu_i^{(2)}) + \log k, \\ R_2: \sum_{i=1}^p \lambda_i x_i &< \sum_{i=1}^p \lambda_i \frac{1}{2} (\mu_i^{(1)} + \mu_i^{(2)}) + \log k. \end{aligned} \quad (17)$$

If we assign *a priori* probabilities, then

$$k = \frac{C(1|2)p_2}{C(2|1)p_1}. \quad (18)$$

In particular, if $k = 1$ (for example, if $C(1|2) = C(2|1)$ and $q_1 = q_2 = \frac{1}{2}$), $\log k = 0$, and the procedure is to compare the discriminant function of the observations with the discriminant function of the averages of the respective means.

If we do not know *a priori* probabilities, we wish to find $\log k = c$, say, so that the expected loss when the observation is from

π_1 is equal to the expected loss when the observation is from π_2 . The probabilities of misclassification can be computed from the distribution of

$$U = \sum_{i=1}^p \lambda_i x_i - \sum_{i=1}^p \lambda_i \frac{1}{2} (\mu_i^{(1)} + \mu_i^{(2)})$$

when x_1, \dots, x_p are from π_1 and when x_1, \dots, x_p are from π_2 . Let a be the "distance" between π_1 and π_2 .

$$a = \sum_{i=1}^p \lambda_i (\mu_i^{(1)} - \mu_i^{(2)}). \quad (19)$$

The distribution of U is normal with the variance a . If the observation is from π_1 the mean of U is $\frac{1}{2}a$; if the observation is from π_2 the mean is $-\frac{1}{2}a$.

The probability of misclassification if the observation is from π_1 is

$$P(2|1, R) = \int_{-\infty}^c \frac{1}{\sqrt{2\pi a}} e^{-\frac{1}{2a}(z-\frac{1}{2}a)^2} dz = \int_{-\infty}^{(c-\frac{1}{2}a)/\sqrt{a}} \frac{1}{\sqrt{2\pi}} e^{-y^2} dy, \quad (20)$$

and the probability of misclassification if the observation is from π_2 is

$$P(1|2, R) = \int_c^{\infty} \frac{1}{\sqrt{2\pi a}} e^{-\frac{1}{2a}(z+\frac{1}{2}a)^2} dz = \int_{(c+\frac{1}{2}a)/\sqrt{a}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2} dy. \quad (21)$$

Figure 2 indicates the two probabilities as the shaded portion in the tails.

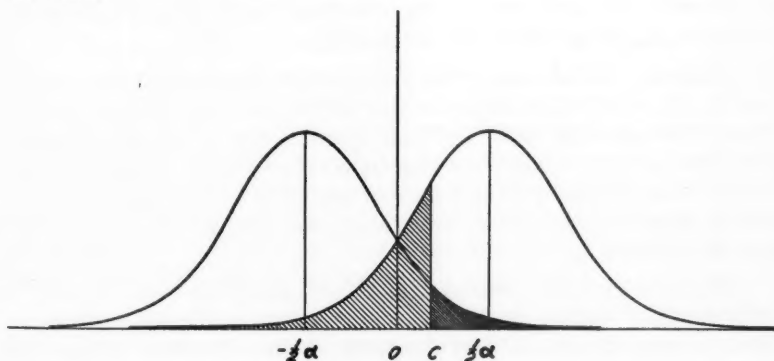


FIGURE 2

We want to choose c so that

$$r(2, R) = C(1|2) \int_{(c+1a)/\sqrt{a}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2} dy = C(2|1) \int_{-\infty}^{(c-1a)/\sqrt{a}} \frac{1}{\sqrt{2\pi}} e^{-y^2} dy = r(1, R). \quad (22)$$

It should be noted that if the costs of misclassification are equal, $c = 0$ and the probability of misclassification is

$$\int_{\sqrt{a}/2}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2} dy. \quad (23)$$

In case the costs of misclassification are unequal, c could be determined to sufficient accuracy by a trial and error method with the normal tables.

In passing let us note that if the set of variances and correlations in one population is not the same as the set in the other population we can still apply the general theory. In this case

$$\log \frac{p_1(x_1, \dots, x_p)}{p_2(x_1, \dots, x_p)}$$

is not a linear function of x_1, \dots, x_p , but a quadratic function. If we have *a priori* probabilities we can give an explicit solution to the problem.

5. Classification into One of Two Multivariate Normal Populations when the Parameters are Estimated.

Thus far we have assumed that the two populations are known exactly. In most applications of this theory the populations are not known, but must be inferred from samples, one from each population. We shall now treat the case in which we have a sample from each of two normal populations and we wish to use that information in classifying another observation as coming from one of the two populations.

Suppose we have a sample $(x_{1\gamma}^{(1)}, \dots, x_{p\gamma}^{(1)})$ ($\gamma = 1, \dots, N^{(1)}$) from π_1 and a sample $(x_{1\gamma}^{(2)}, \dots, x_{p\gamma}^{(2)})$ ($\gamma = 1, \dots, N^{(2)}$) from π_2 . Then we can estimate $\mu_i^{(1)}$ by the mean of the i th variate of the first sample $\bar{x}_i^{(1)}$ and $\mu_i^{(2)}$ by the mean of the second sample $\bar{x}_i^{(2)}$. The

estimate of $\sigma_i \sigma_j \rho_{ij}$ is given by

$$\frac{1}{N^{(1)} + N^{(2)} - 2} \left[\sum_{a=1}^{N^{(1)}} (x_{ia}^{(1)} - \bar{x}_i^{(1)}) (x_{ja}^{(1)} - \bar{x}_j^{(1)}) + \sum_{a=1}^{N^{(2)}} (x_{ia}^{(2)} - \bar{x}_i^{(2)}) (x_{ja}^{(2)} - \bar{x}_j^{(2)}) \right]. \quad (24)$$

We can then substitute these estimates into our definition of U , obtaining a new linear function of x_1, \dots, x_p depending on these estimates. Since there are now sampling variations in the estimates of parameters, we can no longer state that this procedure is best in either of the senses used earlier. However, it seems to be a reasonable procedure.

The exact distributions of this statistic cannot be given explicitly; however, in the Appendix the distribution is indicated as an integral (with respect to three variables). It can be shown that as the sample sizes increase, the distributions of this statistic approach those of the statistic used when the parameters are known. Thus for sufficiently large samples we can proceed exactly as if the parameters were known.

A mnemonic device for the computation of the discriminant function (4) is the introduction of a dummy variate, y , which is equal to a constant (say, 1) when the observation is from π_1 and is equal to another constant (say, 0) when the observation is from π_2 . Then (formally) the regression of this dummy variate y on the observed variates x_1, \dots, x_p over the two samples gives a linear function proportional to the discriminant function. In a sense this linear function is a predictor of the dummy variate y .

Often in studies in psychology or education one has a set of p variates x_1, \dots, x_p and one more variate, y , which is a continuous variate (taking on more than two values). For example, the p variates may be scores on a battery of tests constituting an entrance examination, and the other variable y may be a measure of the degree of success in college (grade average, etc.). We may define "success" then as a score on y equal to or greater than some number a . Given the scores x_1, \dots, x_p for an individual, how should we classify him as to potentially a success or a failure?

Let us assume that x_1, \dots, x_p, y have a joint normal distribution. Then from our theory we can deduce that the proper function of x_1, \dots, x_p to use for classification is the regression of y on

x_1, \dots, x_p , say $\sum_{i=1}^p \beta_i (x_i - \mu_i) + \nu$ where β_i is the regression coefficient of y on x_i , μ_i is the mean of x_i and ν is the mean of y . If this linear function is greater than a constant d we classify the individual as potentially successful, and if this function is less than d we predict failure. The number c depends on the joint normal distribution and the costs of misclassification.

Suppose we do not use this theory but go ahead and find the ordinary discriminant function in this problem by replacing $y \geq a$ by 1 and $y < a$ by 0. Then we arrive at the same linear function. However, the constant $\log k$ will in general be different from the constant d . It would seem that when one must use samples to estimate this linear function that the latter procedure is not as efficient as the former.*

6. Classification into One of Several Groups.

Let us now consider the problem of classifying an observation into one of several groups. Let π_1, \dots, π_m be m populations with density functions $p_1(x_1, \dots, x_p), \dots, p_m(x_1, \dots, x_p)$, respectively. We wish to divide the space of observations into m mutually exclusive and exhaustive regions R_1, \dots, R_m . If an observation falls into R_p we shall say that it comes from π_p . Let the cost of misclassifying an observation from π_g as coming from π_h be $C(h|g)$. The probability of this misclassification is

$$P(h|g, R) = \int_{R_h} p_g(x_1, \dots, x_p) dx_1 \cdots dx_p. \quad (25)$$

If the observation is from π_g , the expected loss or risk is

$$r(g, R) = \sum_{\substack{h=1 \\ h \neq g}}^m C(h|g) P(h|g, R). \quad (26)$$

Suppose we have *a priori* probabilities of the populations, q_1, \dots, q_m . Then the expected loss is

$$\sum_{i=1}^m q_i r(i, R) = \sum_{i=1}^m q_i \left[\sum_{\substack{j=1 \\ j \neq i}}^m C(j|i) P(j|i, R) \right]. \quad (27)$$

We would like to choose R_1, \dots, R_m to make this a minimum.

*Miss Rosedith Sitgreaves is studying this problem at Columbia University.

Since we have *a priori* probabilities for the populations we can define the conditional probability of an observation coming from a population given the values of observed variates, x_1, \dots, x_p . The conditional probability of the observation coming from π_g is

$$\frac{q_g p_g(x_1, \dots, x_p)}{q_1 p_1(x_1, \dots, x_p) + \dots + q_m p_m(x_1, \dots, x_p)}. \quad (28)$$

If we classify the observation as from π_h , the expected loss is

$$\sum_{\substack{g=1 \\ g \neq h}}^m \frac{q_g p_g(x)}{\sum_{k=1}^m q_k p_k(x)} C(h|g), \quad (29)$$

where x denotes x_1, \dots, x_p . We minimize the expected loss at this point if we choose h so as to minimize (29); that is, we consider

$$\sum_{\substack{g=1 \\ g \neq h}}^m q_g p_g(x) C(h|g) \quad (30)$$

for all h and select that h that gives the minimum (if two different indices give the minimum, it is irrelevant which index is selected). This procedure assigns the point x_1, \dots, x_p to one of the R_h . Following this procedure for each point we define our regions R_1, \dots, R_m according to

$$R_k: \sum_{\substack{g=1 \\ g \neq k}}^m q_g p_g(x) C(k|g) < \sum_{\substack{g=1 \\ g \neq h}}^m q_g p_g(x) C(h|g), \quad \begin{matrix} h=1, \dots, m, \\ h \neq k. \end{matrix} \quad (31)$$

If $C(h|g) = 1$ for all g and $h (g \neq h)$, then x_1, \dots, x_p is in R_k if

$$\sum_{\substack{g=1 \\ g \neq k}}^m q_g p_g(x) \leq \sum_{\substack{g=1 \\ g \neq h}}^m q_g p_g(x) \quad (h \neq k). \quad (32)$$

Subtracting $\sum_{\substack{g=1 \\ g \neq k, h}}^m q_g p_g(x)$ from both sides of (32) we obtain

$$R_k: q_h p_h(x) \leq q_k p_k(x) \quad (h \neq k). \quad (33)$$

In this case the point x_1, \dots, x_p is in R_k if k is the index for which $q_k p_k(x)$ is a maximum, that is, π_k is the most probable population.

If we do not have *a priori* probabilities, we cannot define an unconditional expected loss for a classification procedure. Then we consider the maximum of the risk $r(g, R)$ over all values of g . We

would like to choose R_1, \dots, R_m to minimize this maximum expected loss.

It can be shown that the definition of the regions is made by using (31) where q_1, \dots, q_m are replaced by m positive numbers that make

$$r(1, R) = \dots = r(m, R). \quad (34)$$

This number is the expected loss.*

7. Classification into One of Several Multivariate Normal Populations.

As an example of the theory let us consider the case of m multivariate normal populations with the same set of variances and correlations. Let the mean of x_j in π_g be $\mu_j^{(g)}$. Then

$$\log \frac{p_g(x_1, \dots, x_p)}{p_h(x_1, \dots, x_p)} = \sum_{i=1}^p \lambda_{i(g,h)} x_i - \sum_{i=1}^p \lambda_{i(g,h)} \frac{1}{2} (\mu_i^{(g)} + \mu_i^{(h)}), \quad (35)$$

where $\lambda_{1(g,h)}, \dots, \lambda_{p(g,h)}$ are given as the solution to

$$\sum_{j=1}^p \sigma_j \sigma_j \rho_{ij} \lambda_{j(g,h)} = (\mu_i^{(g)} - \mu_i^{(h)}). \quad (36)$$

For the sake of simplicity we assume that the costs of misclassification are equal. If *a priori* probabilities, q_1, \dots, q_m , are known, the regions are defined by

$$R_g: u_{gh}(x_1, \dots, x_p) > \log \frac{q_h}{q_g} = \log q_h - \log q_g, \quad \begin{matrix} h=1, \dots, m, \\ h \neq g, \end{matrix} \quad (37)$$

where $u_{gh}(x_1, \dots, x_p)$ is (35).

If *a priori* probabilities are not known, the regions are defined by

$$R_g: u_{gh}(x_1, \dots, x_p) > c_h - c_g, \quad \begin{matrix} h=1, \dots, m, \\ h \neq g, \end{matrix} \quad (38)$$

where the numbers c_1, \dots, c_m are to be determined so that

$$P(1|1, R) = \dots = P(m|m, R). \quad (39)$$

To determine these constants we use the fact that if the observation is from π_g , $u_{gh}(x_1, \dots, x_p)$, $h=1, \dots, m$ and $h \neq g$, have a joint normal distribution with means

*The theory was first given for the case of equal costs of misclassification by R. von Mises (7).

$$Eu_{gh}(x_1, \dots, x_p) = \frac{1}{2} \sum_{i=1}^p \lambda_i^{(g,h)} (\mu_i^{(g)} - \mu_i^{(h)}). \quad (40)$$

The variance of $u_{gh}(x_1, \dots, x_p)$ is twice (40), and the covariance between $u_{gh}(x_1, \dots, x_p)$ and $u_{gk}(x_1, \dots, x_p)$ is

$$\sum_{i=1}^p \lambda_i^{(g,h)} (\mu_i^{(g)} - \mu_i^{(k)}) = \sum_{i=1}^p \lambda_i^{(g,k)} (\mu_i^{(g)} - \mu_i^{(h)}). \quad (41)$$

From this one can determine $P(g|g,R)$ for any set of constants c_1, \dots, c_m .

In any given case it would be exceedingly difficult to determine these constants so that (39) would be satisfied. Perhaps a reasonable practical procedure would be to let $c_1 = \dots = c_m = 1/m$ and then see whether (39) were approximately satisfied. However, even the computation of $P(g|g,R)$ for given constants is usually far from easy.

It should be pointed out that this procedure divides our space by means of hyperplanes. If $p = 2$ and $m = 3$, the division is by half-lines as in Figure 3.

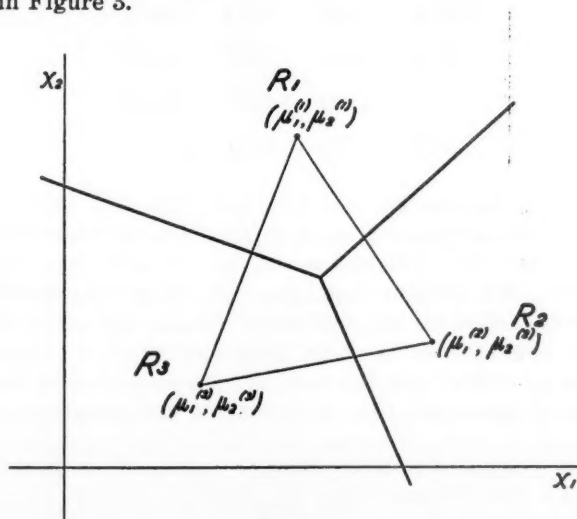


FIGURE 3

Finally, we remark that if the populations are unknown, we can estimate the parameters by means of samples, one from each population. If the samples are large enough, the above procedures can be used as if the parameters were known.

8. *An Example of Classification into One of Several Multivariate Normal Populations.*

In (6) Rao considers three populations consisting of the Brahmin caste (π_1), the Artisan caste (π_2), and the Korwa caste (π_3) of India. The measurements for each individual of a caste are stature (x_1), sitting height (x_2), nasal depth (x_3), and nasal height (x_4). The means of these variables in the three populations are

	Brahmin (π_1)	Artisan (π_2)	Korwa (π_3)
Stature (x_1)	164.51	160.53	158.17
Sitting height (x_2)	86.43	81.47	81.16
Nasal depth (x_3)	25.49	23.84	21.44
Nasal height (x_4)	51.24	48.62	46.72

The matrix of correlations for all the populations is

1.0000	.5849	.1774	.1974
.5849	1.0000	.2094	.2170
.1774	.2094	1.0000	.2910
.1974	.2170	.2910	1.0000

The standard deviations are $\sigma_1 = 5.74$, $\sigma_2 = 3.20$, $\sigma_3 = 1.75$, $\sigma_4 = 3.50$. We assume that each population is normal. Our problem is to divide the space of the four variables x_1, x_2, x_3, x_4 into three regions of classification. We assume that the costs of misclassifications are equal. We shall find (i) a set of regions under the assumption that drawing a new observation from each population is equally likely ($q_1 = q_2 = q_3 = 1/3$) and (ii) a set of regions such that the largest probability of misclassification is minimized (the minimax solution).

We first compute the coefficients $\lambda_j^{(1,2)}$ and $\lambda_j^{(1,3)}$ defined by (36); the other λ 's are obtained from the relations $\lambda_j^{(g,h)} = -\lambda_j^{(h,g)}$ and $\lambda_j^{(2,3)} = \lambda_j^{(1,3)} - \lambda_j^{(1,2)}$. After calculating $\sum_{i=1}^4 \lambda_i^{(g,h)} \frac{1}{2}(\mu_i^{(g)} + \mu_i^{(h)})$, we obtain the "discriminant functions" given by* (35).

*Due to an error in computations Rao's discriminant functions (6) are incorrect. I am indebted to Mr. Peter Frank for assistance in the computations.

$$\begin{aligned}u_{12}(x_1, x_2, x_3, x_4) &= -.0708x_1 + .4990x_2 + .3373x_3 + .0887x_4 + 43.13, \\u_{13}(x_1, x_2, x_3, x_4) &= .0003x_1 + .3550x_2 + 1.1063x_3 + .1375x_4 + 62.49, \\u_{23}(x_1, x_2, x_3, x_4) &= .0711x_1 - .1440x_2 + .7690x_3 + .0488x_4 + 19.36.\end{aligned}$$

The other three functions are $u_{21}(x_1, x_2, x_3, x_4) = -u_{12}(x_1, x_2, x_3, x_4)$, $u_{31}(x_1, x_2, x_3, x_4) = -u_{13}(x_1, x_2, x_3, x_4)$, and $u_{32}(x_1, x_2, x_3, x_4) = -u_{23}(x_1, x_2, x_3, x_4)$. If there are *a priori* probabilities and they are equal, the best set of regions of classification are $R_1: u_{12}(x_1, x_2, x_3, x_4) \geq 0$, $u_{13}(x_1, x_2, x_3, x_4) \geq 0$; $R_2: u_{21}(x_1, x_2, x_3, x_4) \geq 0$, $u_{23}(x_1, x_2, x_3, x_4) \geq 0$; and $R_3: u_{31}(x_1, x_2, x_3, x_4) \geq 0$, $u_{32}(x_1, x_2, x_3, x_4) \geq 0$. For example, if we obtain an individual with measurements x'_1, x'_2, x'_3, x'_4 such that $u_{12}(x'_1, x'_2, x'_3, x'_4) \geq 0$ and $u_{13}(x'_1, x'_2, x'_3, x'_4) \geq 0$ we classify him as a Brahmin.

To find the probabilities of misclassification when an individual is drawn from population π_j we need the means, variances, and covariance of the proper pairs of u 's. They are

Population of x_1, x_2, x_3, x_4	u 's	Means	Standard Deviation	Correlation
π_1	u_{12}	1.491	1.727	.8658
	u_{13}	3.487	2.641	
π_2	u_{21}	1.491	1.727	.3894
	u_{23}	1.031	1.436	
π_3	u_{31}	3.487	2.641	.7983
	u_{32}	1.491	1.436	

The probabilities of misclassification are then obtained by use of the tables for the bivariate normal distribution (5). These probabilities are .21 for π_1 , .42 for π_2 , and .25 for π_3 . For example, if measurements are made on a Brahmin, the probability that he is classified as an Artisan or Korwa is .21.

The minimax solution is obtained by finding the constants c_1 , c_2 , and c_3 for (38) so that the probabilities of misclassification are equal. The regions of classification are

$$\begin{aligned}R'_1: u_{12}(x_1, x_2, x_3, x_4) &\geq .54, & u_{13}(x_1, x_2, x_3, x_4) &\geq .39; \\R'_2: u_{21}(x_1, x_2, x_3, x_4) &\geq -.54, & u_{23}(x_1, x_2, x_3, x_4) &\geq -.25; \\R'_3: u_{31}(x_1, x_2, x_3, x_4) &\geq -.39, & u_{32}(x_1, x_2, x_3, x_4) &\geq .25.\end{aligned}$$

The common probability of misclassification (to two decimal places)

is .30. Thus, the maximum probability of misclassification has been reduced from .42 to .30.

Appendix: On the Distribution of the Classification Statistic.

In (8) Wald considered the distributions of a class of statistics of which the classification statistic given in Section 5 is a special case. Wald showed that such a statistic can be written as a function of three quantities, say, m_1, m_2, m_3 . His expression for the distribution of m_1, m_2, m_3 involved an expected value that was not evaluated. In this appendix we shall give this distribution of m_1, m_2 , and m_3 explicitly.

Let the elements defined by (24) be s_{ij} and let $\|s_{ij}\|^{-1} = \|s^{ij}\|$. Then the sample estimate of λ_i is $\sum_{j=1}^p s^{ij}(\bar{x}_j^{(1)} - \bar{x}_j^{(2)})$, and the sample equivalent of U is

$$W = \sum_{i,j=1}^p [x_i - \frac{1}{2}(\bar{x}_i^{(1)} + \bar{x}_i^{(2)})] s^{ij}(\bar{x}_j^{(1)} - \bar{x}_j^{(2)}). \quad (42)$$

Let

$$t_{i,n+1} = \sqrt{(N^{(1)} + N^{(2)}) / (N^{(1)} + N^{(2)} + 1)} (x_i - \bar{x}_i),$$

$$t_{i,n+2} = \sqrt{(N^{(1)}N^{(2)}) / (N^{(1)} + N^{(2)})} (\bar{x}_i^{(1)} - \bar{x}_i^{(2)}),$$

where

$$\bar{x}_i = (N^{(1)}\bar{x}_i^{(1)} + N^{(2)}\bar{x}_i^{(2)}) / (N^{(1)} + N^{(2)}) \text{ and } n = N^{(1)} + N^{(2)} - 2.$$

Then

$$W = \sqrt{(N^{(1)} + N^{(2)} + 1) / N^{(1)}N^{(2)}} W_1 + \sqrt{(N^{(1)} - N^{(2)}) / 2N^{(1)}N^{(2)}} W_2, \quad (43)$$

where

$$W_1 = \sum_{i,j=1}^p t_{i,n+1} s^{ij} t_{j,n+2}, \quad (44)$$

$$W_2 = \sum_{i,j=1}^p t_{i,n+2} s^{ij} t_{j,n+2}.$$

W_2 is proportional to Hotelling's generalized T^2 statistic for testing $\mu_i^{(1)} = \mu_i^{(2)}$.

The set $t_{i,n+1}$ and the set $t_{i,n+2}$ have multivariate normal distributions with variances σ_i^2 and correlations ρ_{ij} . The set $t_{i,n+1}, t_{i,n+2}$ and s^{ij} have the joint distribution assumed by Wald, and $\mathcal{E}t_{i,n+1}$ is proportional to $\mathcal{E}t_{i,n+2}$. Wald expresses W_1 as

$$W_1 = n \frac{m_3}{(1-m_1)(1-m_2) - m_3^2}, \quad (45)$$

where m_1 , m_2 , and m_3 are certain functions of s_{ij} , $t_{i,n+1}$, $t_{i,n+2}$. It is easy to verify that

$$W_2 = n \frac{m_1 - m_1 m_2 + m_3^2}{(1-m_1)(1-m_2) - m_3^2}.$$

The joint density of m_1 , m_2 , and m_3 involves the expected value of the determinant $|\sum_{v=1}^p t_{iv} t_{jv}|$, where t_{iv} have a certain normal distribution. This expected value which Wald did not evaluate is a special case* of (1). Thus we obtain as the density of m_1 , m_2 , and m_3

$$K e^{-\frac{1}{2} \alpha^2 (c^2 + k^2)} (m_1 m_2 - m_3^2)^{\frac{1}{2}(p-3)} [(1-m_1)(1-m_2) - m_3^2]^{\frac{1}{2}(n-p-1)} \\ \cdot \sum_{v=0}^{\infty} \frac{\Gamma(\frac{1}{2}n + v + 1) \alpha^v (k^2 m_1 + 2k c m_3 + c^2 m_2)^v}{\Gamma(\frac{1}{2}p + v) v! 2^v}, \quad (47)$$

$$(m_1 m_2 - m_3^2 \geq 0, (1-m_1)(1-m_2) - m_3^2 \geq 0, 0 \leq m_1, m_2 \leq 1),$$

where

$$k = \sqrt{N^{(1)} N^{(2)} / (N^{(1)} + N^{(2)})},$$

$$c = N^{(2)} / \sqrt{(N^{(1)} + N^{(2)}) (N^{(1)} + N^{(2)} + 1)}$$

if the observation is from π_1 , and

$$c = -N^{(1)} / \sqrt{(N^{(1)} + N^{(2)}) (N^{(1)} + N^{(2)} + 1)}$$

if the observation is from π_2 , and K is a number chosen to make the integral of (47) over the entire range of m_1 , m_2 , and m_3 equal to one (α is defined in Section 5). In principle the density of W can be obtained from (47) by integrating out two variables imposing (43), (45), and (46).

*One needs to justify the expression in (1) for the non-central Wishart distribution when v runs over the same range as i . This can be done by deriving the non-central Wishart distribution from the distribution of correlation coefficients, χ^2 -distributions, and a non-central χ^2 -distribution.

REFERENCES

1. Anderson, T. W. The non-central Wishart distribution and certain problems of multivariate statistics. *Ann. math. Stat.*, 1946, 17, 409-431.
2. Brown, George W. Basic principles for construction and application of discriminators. *J. clin. Psychol.*, 1950, 6, 58-61.
3. Fisher, R. A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, 1936, 7, 179-188.
4. Fisher, R. A. The statistical utilization of multiple measurements. *Ann. Eugen.*, 1938, 8, 376-386.
5. Pearson, K. Tables for Statisticians and Biometricians, Part II. Cambridge University Press, 1931.
6. Rao, C. R. The utilization of multiple measurements in problems of biological classification. *J. roy. stat. Soc., Series B*, 1948, 10, 159-193.
7. von Mises, R. On the classification of observation data into distinct groups. *Ann. math. Stat.*, 1945, 16, 68-73.
8. Wald, A. On a statistical problem arising in the classification of an individual into one of two groups. *Ann. math. Stat.*, 1944, 15, 145-163.
9. Wald, A. Foundations of a general theory of sequential decision functions. *Econometrica*, 1947, 15, 279-313.
10. Welch, B. L. Note on discriminant functions. *Biometrika*, 1939, 13, 218-220.

Manuscript received 8/14/50

Revised manuscript received 11/27/50

THE RELATIONSHIP BETWEEN THE METHOD OF SUCCESSIVE RESIDUALS AND THE METHOD OF EXHAUSTION

WILBUR L. LAYTON
UNIVERSITY OF MINNESOTA

The relationship between Horst's method of successive residuals and Gengerelli's method of exhaustion is demonstrated by transforming both methods into L notation. The L notation form is much more efficient computationally.

Horst (2), in 1934, presented the development of a method of item analysis called the method of successive residuals. Recently, Gengerelli (1) presented an "exhaustion" method for calculating regression coefficients. The present article proposes to show the relationship between the two methods by transforming them both into L notation (3), computationally more efficient.

In the method of successive residuals that item having the largest coefficient of correlation with the criterion is selected as the first item of a composite. That portion of the criterion which is predicted by the first selected item is subtracted from the criterion. This leaves a criterion residual. The next item selected is that item which has the highest coefficient of correlation with the criterion residual. That part of the criterion residual which is predicted by the second selected item is subtracted from the criterion residual. This leaves a second criterion residual. The third item selected is that item which has the highest coefficient of correlation with the second criterion residual. This process is continued in the same manner to select the n items to be included in the composite.

Horst presents the general formula for a criterion residual as

$$C_{e-1}(A_e X_e + B_e) = C_e \quad (1)$$

where A_e and B_e have been determined by the method of least squares and C_e indicates that part of the criterion which cannot be predicted by item e , i.e., C_e is a criterion residual.

The correlation between the criterion residual C_e and the next item to be selected is given by Horst as

$$r_{C_e X_k} = r_{(C_{e-1} - A_e X_e) X_k} \quad (2)$$

where k is the item being considered for inclusion in the composite. His computational formula is

$$N \sigma_{C_e} r_{(C_{e-1} - A_e X_e) X_k} = \frac{N \sum (C_{e-1} - A_e X_e) X_k - \sum (C_{e-1} - A_e X_e) N_k}{\sqrt{N N_k - N_k^2}} \quad (3)$$

where N_k is the number of individuals who marked item k in a specified manner and is consequently $\sum X_k$ and $\sum X_k^2$.

The gross score weight, A_e , is defined by Horst as

$$A_e = \frac{N \sigma_{C_{e-1}} r_{C_{e-1} X_e}}{\sqrt{N N_e - N_e^2}} \quad (4)$$

These two formulas provide a method for assigning differential gross score weights to items.

Horst extended his method to provide a means for assigning positive or negative unit weights to items. In so doing, he defined the coefficient of correlation between successive criterion residuals and the successive items to be selected in terms of the original criterion rather than in terms of the criterion residuals themselves. This formula is

$$\frac{N \sigma_{C_e} r_{C_e X_k} = N \sum [C_0 - A_e (\pm N_{1k, \dots, z} N_{ek})] X_k - N_k [\sum C_0 - A_e (\pm N_{1, \dots, z} N_e)]}{\sqrt{N N_k - N_k^2}} \quad (5)$$

where N_{ek} is the number of people answering both items e and k in a specified manner, and is therefore $\sum X_e X_k$. By analogy we can write a similar formula in which differential item weights are involved as

$$\frac{N \sigma_{C_e} r_{C_e X_k} = N \sum [C_0 - (A_1 X_1 + \dots + A_e X_e)] X_k - \sum X_k \sum [C_0 - (A_1 X_1 + \dots + A_e X_e)]}{\sqrt{N \sum X_k^2 - (\sum X_k)^2}} \quad (6)$$

It is obvious that formula (6) can be used to select tests for batteries as well as selecting items for tests.

Formula (6) can be rewritten as

$$\frac{N \sigma_{C_e} r_{C_e X_k} = N \sum C_0 X_k - N A_1 \sum X_1 X_k - \dots - N A_e \sum X_e X_k - \sum X_k \sum C_0 + A_1 \sum X_1 \sum X_k + \dots + A_e \sum X_e \sum X_k}{\sqrt{N \sum X_k^2 - (\sum X_k)^2}} \quad (7)$$

Collecting terms, we find

$$\frac{N \sigma_{C_e} r_{C_e X_k} = N \sum C_0 X_k - \sum X_k \sum C_0 - A_1 (N \sum X_1 X_k - \sum X_1 \sum X_k) - \dots - A_e (N \sum X_e X_k - \sum X_e \sum X_k)}{\sqrt{N \sum X_k^2 - (\sum X_k)^2}} \quad (8)$$

Transforming into L notation after Toops (3), we have

$$N \sigma_{C_e} r_{C_e X_k} = \frac{L_{0k} - A_1 L_{1k} - \dots - A_e L_{ek}}{\sqrt{L_{kk}}}, \quad (9)$$

where

$$L_{ii} = N \sum X_i^2 - (\sum X_i)^2 \text{ and } L_{ij} = N \sum X_i X_j - \sum X_i \sum X_j.$$

It is clear that dividing (9) by $\sqrt{L_{kk}}$ we obtain the formula for the differential weight A_k for item k as

$$A_k = \frac{L_{0k} - A_1 L_{1k} - \dots - A_e L_{ek}}{L_{kk}}. \quad (10)$$

Formula (9) can be considered a general formula to be used in selecting successive items or tests and formula (10) as a general formula for determining the differential weight for an item or test after it has been selected.

Let us now turn to Gengerelli's method of exhaustion. The method of exhaustion is used to determine successively the values of the regression coefficients by solving in turn Gengerelli's formulas of the types

$$\beta_1 = r_{01}, \quad (11)$$

$$\beta_2 = r_{02} - \beta_1 r_{12}, \text{ and} \quad (12)$$

$$\beta_3 = r_{03} - \beta_1 r_{13} - \beta_2 r_{23}. \quad (13)$$

Substituting (11) in (12) we have

$$\beta_2 = r_{02} - r_{01}r_{12}, \quad (14)$$

and substituting (11) and (14) in (13) we have

$$\beta_3 = r_{03} - r_{01}r_{13} - (r_{02} - r_{01}r_{12})r_{23}. \quad (15)$$

Transforming β_1 , β_2 and β_3 into gross score weights from formulas (11), (14), and (15) we have

$$b_1 = r_{01} \frac{\sigma_0}{\sigma_1}, \quad (16)$$

$$b_2 = r_{02} \frac{\sigma_0}{\sigma_2} - r_{01}r_{12} \frac{\sigma_0}{\sigma_2}, \quad (17)$$

and

$$b_3 = r_{03} \frac{\sigma_0}{\sigma_3} - r_{01}r_{13} \frac{\sigma_0}{\sigma_3} - (r_{02} - r_{01}r_{12})r_{23} \frac{\sigma_0}{\sigma_3}. \quad (18)$$

Now

$$r_{xy} = \frac{L_{xy}}{\sqrt{L_{yy}} \sqrt{L_{xx}}}, \quad (19)$$

and

$$\sigma_x = \sqrt{\frac{L_{xx}}{N^2}}. \quad (20)$$

Therefore

$$b_1 = \frac{L_{01}}{\sqrt{L_{00}} \sqrt{L_{11}}} \frac{\sqrt{\frac{L_{00}}{N^2}}}{\sqrt{\frac{L_{11}}{N^2}}}, \quad (21)$$

$$b_2 = \frac{L_{02}}{\sqrt{L_{00}} \sqrt{L_{22}}} \frac{\sqrt{\frac{L_{00}}{N^2}}}{\sqrt{\frac{L_{22}}{N^2}}} - \frac{L_{01}}{\sqrt{L_{00}} \sqrt{L_{11}}} \frac{L_{12}}{\sqrt{L_{11}} \sqrt{L_{22}}} \frac{\sqrt{\frac{L_{00}}{N^2}}}{\sqrt{\frac{L_{22}}{N^2}}}, \quad (22)$$

and

$$b_3 = \frac{L_{03}}{\sqrt{L_{00}} \sqrt{L_{33}}} \cdot \frac{\sqrt{\frac{L_{00}}{N^2}}}{\sqrt{\frac{L_{33}}{N^2}}} - \frac{L_{01}}{\sqrt{L_{01}} \sqrt{L_{11}}} \cdot \frac{L_{13}}{\sqrt{L_{11}} \sqrt{L_{33}}} \cdot \frac{\sqrt{\frac{L_{00}}{N^2}}}{\sqrt{\frac{L_{33}}{N^2}}} \quad (23)$$

$$- \left(\frac{L_{02}}{\sqrt{L_{00}} \sqrt{L_{22}}} - \frac{L_{01}}{\sqrt{L_{11}} \sqrt{L_{11}}} \cdot \frac{L_{12}}{\sqrt{L_{11}} \sqrt{L_{22}}} \right) \frac{L_{23}}{\sqrt{L_{22}} \sqrt{L_{33}}} \cdot \frac{\sqrt{\frac{L_{00}}{N^2}}}{\sqrt{\frac{L_{33}}{N^2}}}.$$

Simplifying

$$b_1 = \frac{L_{01}}{L_{11}}, \quad (24)$$

$$b_2 = \frac{L_{02} - \frac{L_{01}}{L_{11}} \cdot L_{12}}{L_{22}}, \quad (25)$$

and

$$b_3 = \frac{L_{03} - \frac{L_{01}}{L_{11}} \cdot L_{13} - \frac{L_{02}}{L_{22}} \cdot \frac{L_{01}}{L_{11}} - \frac{L_{12}}{L_{22}} \cdot L_{23}}{L_{33}}. \quad (26)$$

Then:

$$b_2 = \frac{L_{02} - b_1 L_{12}}{L_{22}}, \quad (27)$$

and

$$b_3 = \frac{L_{03} - b_1 L_{13} - b_2 L_{23}}{L_{33}}. \quad (28)$$

In general,

$$b_m = \frac{L_{0n} - b_1 L_{1n} - \dots - b_{n-1} L_{n-1n}}{L_{nn}}. \quad (29)$$

It is apparent that formula (29) based on Gengerelli's method of exhaustion is identical with formula (10) based on an extension of Horst's method of successive residuals.

Let us now return to formulas (9)

$$N \sigma_{c_e} r_{c_e x_k} = \frac{L_{0k} - A_1 L_{1k} - \dots - A_e L_{ek}}{\sqrt{L_{kk}}},$$

and (10)

$$A_k = \frac{L_{0k} - A_1 L_{1k} - \dots - A_e L_{ek}}{L_{kk}}.$$

In L notation, the formula for the coefficient of correlation between the criterion and the composite of selected tests or items becomes

$$\frac{r_{c_0(A_1 X_1 + A_2 X_2 + \dots + A_e X_e)}}{A_1 L_{01} + A_2 L_{02} + \dots + A_e L_{0e}} = \frac{\sqrt{L_{00}} \sqrt{A_1^2 L_{11} + A_2^2 L_{22} + \dots + A_e^2 L_{ee} + 2(\sum A_i A_j L_{ij})}}{\quad} \quad (30)$$

It will be noted that the numerator increases accumulatively from selection to selection by the amount $A_e L_{0e}$ and that the right hand element of the denominator increases accumulatively by the amount $A_e^2 L_{ee} + 2\sum A_e A_j L_{ej}$. It is obvious that formula (10) is merely formula (9) divided by $\sqrt{L_{kk}}$. Formula (9) makes for easy computation since the numerator accumulates from each item or test selected to the next.

Further, if one is dealing with large numbers of items or tests the L_{ij} of a specific item or test with all the other variables being considered need not be computed until the variable has definitely been selected as part of the composite. Punched card methods aid in this since the sums necessary for the computation of the L 's are quickly run off on an IBM tabulator and summary punch. If one is dealing with test items, the IBM sorter will deliver the $\sum X_i X_j$ of the selected item with other items from multiple-punched cards very rapidly. From the obtained sums one can compute L 's at a fast rate using a calculator which has automatic positive and negative multiplication.

REFERENCES

1. Gengerelli, J. A. A simplified method of approximating multiple regression coefficients. *Psychometrika*, 1948, 13, 135-146.
2. Horst, Paul. Item analysis by the method of successive residuals. *Journal of Experimental Education*, 1934, 2, 254-263.
3. Toops, H. A. The L-Method. *Psychometrika*, 1941, 6, 249-266.

Manuscript received 4/24/50

THE RELATIONSHIP BETWEEN THE VALIDITY OF A SINGLE TEST AND ITS CONTRIBUTION TO THE PREDICTIVE EFFICIENCY OF A TEST BATTERY

PAUL HORST

EDUCATIONAL TESTING SERVICE

Typical selection or classification testing programs should provide for improvement of the predictive efficiency of the test battery. Such provision calls for the administration of experimental tests along with the operational battery administration and follow-up analysis to determine the value of the experimental material. It is possible to determine without waiting for criterion data what the validity of the experimental test must be in order to improve the battery validity. The method together with the proof is presented.

I. The Method

Any serious program concerned with the prediction of criterion measures whether in industry, education, the military services, or elsewhere will ordinarily utilize a battery of prediction measures rather than a single measure. Furthermore it is urged that for a given criterion, the scores on the prediction measures be combined into a single predicted criterion score by means of "least square" regression weights. An adequate testing program should also incorporate within its operational and administrative framework provision for the improvement of the prediction battery in terms of its multiple correlation with the criterion. The most obvious method for improving the battery is to add a test which will improve the multiple correlation. But to find whether a test will improve the multiple correlation, it is necessary to have its correlations with all the other tests in the battery and also its correlation with the criterion. In many cases the collection of criterion data and its collation with test data is a costly and time-consuming process. Often it is necessary to wait for months or even years after test data are obtained before criterion data on the same cases become available. This is true when tests are given to candidates for training or applicants for jobs in which success cannot be adequately determined until after months or years of performance.

In such cases it may be very useful to know soon after scores on the experimental test become available what the validity of the test would have to be in order for the test to make a specified increase in battery validity. If it turns out that the validity must be a value which is highly improbable of attainment, then one may better forego the time and effort of collecting validity data later on and proceed at once to the development and tryout of other test material. Fortunately, if we know the correlation of an experimental test with each test in the predictive battery, it is possible to determine what its validity must be in order to increase the battery validity by a specified amount.

The formula for determining this required validity coefficient is given by:

$$r_{ck} = r_{ks}R_{c,t} \pm \sqrt{a(a + 2R_{c,t})(1 - R_{k,t}^2)}, \quad (A)$$

where

- a = the specified increase in the multiple correlation,
- r_{ck} = the validity required of test k to achieve the increase a ,
- $R_{c,t}$ = the multiple correlation of the test battery, excluding test k , with the criterion,
- r_{ks} = the correlation of test k with the predicted criterion score derived from the test battery with test k excluded, and
- $R_{k,t}$ = the multiple correlation of test k with the test battery.

Formula A requires the multiple correlation of the new test with the battery. It also requires its correlation with the predicted criterion scores. If we are willing to take the latter as a sufficiently close estimate of the former, we can write Formula A thus:

$$r_{ck} = r_{ks}R_{c,t} \pm \sqrt{a(a + 2R_{c,t})(1 - r_{ks}^2)}. \quad (B)$$

It will be noted that Formulas A and B yield two values for r_{ck} , one for the plus sign and one for the minus sign. It will also be noted that, if r_{ck} has the value indicated by the minus sign, test k will have a negative regression weight, while if r_{ck} has the value indicated by the plus sign, the regression weight will be positive. In the former case the test will serve as a suppression test.

The value of $R_{k,t}$ will, in general, be greater than that of r_{ks} . For any specified increase a in the multiple correlation, Formula B will, therefore, overestimate the value required of r_{ck} if the plus sign is used and underestimate it if the minus sign is used. Formula B,

however, is not very sensitive to variations of r_{ks} under the radical. For example, let us assume that $R_{c,t} = .50$, $r_{ks} = .40$, and $a = .02$. The required validity, r_{ck} , of test k would then be .331 for the plus sign and .069 for the minus sign. Suppose now, that instead of using r_{ks} , or .40, in the radical, we had used the correct value, $R_{k,t}$, the multiple correlation of the test with the battery. This value would, in general, be greater than .40. Assuming it to be .50, Formula A would give .324 as the required validity of test k for the plus sign and .076 for the minus sign. These values differ by less than one in the second decimal place from the approximate values given by Formula B.

In practice, Formula B is much easier to use than Formula A, since it requires neither the correlation of the experimental test with each test in the battery nor the multiple correlation of the test with the battery. Formula B is the equivalent of treating the existing battery as one variable and the new test as a second variable. That is, we ask what the validity of a new test must be if it is to make a specific addition to the present battery as it is currently weighted, knowing the correlation of the new test with the battery. Formula A differs from Formula B in that it permits the present weights of tests in the battery to be changed to the actual regression weights, including the new test, whereas Formula B uses the existing weights for tests already in the battery.

To illustrate the use of the formula let us assume that we have a battery of tests whose multiple correlation with the criterion is .50. This value is fairly typical of the multiple correlations between test batteries and a wide variety of academic and industrial criteria. Substituting .50 for $R_{c,t}$ in Formula B we have

$$r_{ck} = .5r_{ks} \pm \sqrt{a(a+1)(1-r_{ks}^2)}. \quad (C)$$

Table 1, based on Formula C, has been prepared to indicate for various values of a and r_{ks} the required value of r_{ck} . For example, suppose the experimental test correlates .35 with predicted criterion score and that, in order to justify its inclusion in the battery, it is specified that it shall increase the multiple correlation of .50 by at least .02. In the r_{kt} column at the left we find the value, .35; and, following this row to the column headed ".02," we read the values, .309 and .041. Therefore, if a test correlates .35 with the predicted criterion score and is to increase the multiple correlation by .02, its validity must lie either above or below the range, .041 to .309. Similarly, for a predicted criterion correlation of .45 and an increase of

.03 in the multiple correlation, the validity must be above or below the range, .068 to .382.

TABLE 1

The Validity Required of Test k in Terms of its Correlation with Predicted Criterion Score and of Desired Increase in Multiple Correlation

r_{ks}	Validity Required if R of .50 is to be increased by					
	.01	.02	.03	.04	.05	
.05	.125	.168	.201	.229	.254	+
	-.075	-.118	-.151	-.179	-.204	-
.10	.150	.192	.225	.253	.278	+
	-.050	-.092	-.125	-.153	-.178	-
.15	.174	.216	.249	.277	.301	+
	-.024	-.066	-.099	-.127	-.151	-
.20	.198	.240	.272	.300	.324	+
	.002	-.040	-.072	-.100	-.124	-
.25	.222	.263	.295	.322	.347	+
	.028	-.013	-.045	-.072	-.097	-
.30	.245	.286	.318	.345	.368	+
	.055	.014	-.018	-.045	-.068	-
.35	.269	.309	.340	.366	.390	+
	.081	.041	.010	-.016	-.040	-
.40	.292	.331	.361	.387	.410	+
	.108	.069	.039	.013	-.010	-
.45	.314	.353	.382	.407	.429	+
	.136	.097	.068	.043	.021	-
.50	.337	.374	.402	.427	.448	+
	.163	.126	.098	.073	.052	-
.55	.358	.394	.422	.445	.466	+
	.192	.156	.128	.105	.084	-
.60	.380	.414	.441	.463	.483	+
	.220	.186	.159	.137	.117	-
.65	.401	.434	.459	.480	.499	+
	.249	.216	.191	.170	.151	-
.70	.421	.452	.476	.496	.514	+
	.279	.248	.224	.204	.186	-

It should be pointed out that a test might be considered for the battery even though it does not increase the predictive value of the

current battery. For example, the new test may be simpler in construction or require less testing time than one or more of the tests already in the battery. It might then be possible to substitute the new test for such tests without lowering the predictive value of the test battery. Thus, economy of time and effort would be achieved without loss of predictive efficiency.

II. *Proof of the Method*

The proof of Formula A may be developed as follows:

Let

- r = the matrix of intercorrelations of the tests in the battery exclusive of test k ,
- ρ = the matrix of intercorrelations of the tests in the battery including test k ,
- r_c = the vector of correlations of the tests in the battery, exclusive of test k , with the criterion,
- ρ_c = the vector of correlations of the tests in the battery, including test k , with the criterion,
- r_k = the vector of correlations of test k with the other tests in the battery,
- β = the vector of regression weights of the tests in the battery, including test k , with the criterion,
- β_c = the vector of regression weights of the tests in the battery, exclusive of test k , with the criterion,
- β_k = the vector of regression weights of the tests in the battery with test k ,
- r_{ck} = the correlation of test k with the criterion,
- $R_{c,t}$ = the multiple correlation of the tests in the battery, exclusive of test k , with the criterion,
- $R_{c,T}$ = the multiple correlation of the tests in the battery, including test k , with the criterion, and
- $R_{k,t}$ = the multiple correlation of test k with the other tests in the battery.

From the above definitions, we have the relations

$$\rho = \begin{vmatrix} r & r_k \\ r'_k & 1 \end{vmatrix} \quad (1)$$

$$\rho_c = \begin{vmatrix} r_c \\ r_{ck} \end{vmatrix} \quad (2)$$

From the definitions and well-known formulas, we have

$$\rho^{-1}\rho_c = \beta. \quad (3)$$

$$\rho'_c \rho^{-1} \rho_c = R^2_{c.T}. \quad (4)$$

$$\rho'_c \beta = R^2_{c.T}. \quad (5)$$

$$r^{-1}r_c = \beta_c. \quad (6)$$

$$r'_c r^{-1} r_c = R^2_{c.t}. \quad (7)$$

$$r'_c \beta_c = R^2_{c.t}. \quad (8)$$

$$r^{-1}r_k = \beta_k. \quad (9)$$

$$r'_k r^{-1} r_k = R^2_{k.t}. \quad (10)$$

$$r'_k \beta_k = R^2_{k.t}. \quad (11)$$

From (1) we have

$$\rho^{-1} = \begin{vmatrix} r & r_k \\ r'_k & 1 \end{vmatrix}^{-1}. \quad (12)$$

To evaluate the right-hand side of (12), we write

$$\begin{vmatrix} r & r_k \\ r'_k & 1 \end{vmatrix}^{-1} = \begin{vmatrix} s & u \\ u' & d \end{vmatrix}. \quad (13)$$

From (13) we have

$$\begin{vmatrix} r & r_k \\ r'_k & 1 \end{vmatrix} \begin{vmatrix} s & u \\ u' & d \end{vmatrix} = \begin{vmatrix} I & 0 \\ 0' & 1 \end{vmatrix}. \quad (14)$$

Expanding the left-hand side of (14) gives

$$rs + r_k u' = I. \quad (15)$$

$$ru + r_k d = 0. \quad (16)$$

$$r'_k s + u' = 0'. \quad (17)$$

$$r'_k u + d = 1. \quad (18)$$

Premultiplying (15) by r^{-1} and transposing,

$$s = r^{-1} - r^{-1}r_k u'. \quad (19)$$

Substituting (19) in (17) and solving for u ,

$$u = \frac{-r^{-1}r_k}{1 - r'_k r^{-1}r_k}. \quad (20)$$

Substituting (20) in (19) gives

$$s = r^{-1} + \frac{r^{-1}r_k r'_k r^{-1}}{1 - r'_k r^{-1}r_k}. \quad (21)$$

Substituting (20) in (18) and solving for d ,

$$d = \frac{1}{1 - r'_k r^{-1}r_k}. \quad (22)$$

Substituting (9) and (10) in (21), (20), and (22), respectively,

$$s = r^{-1} + \frac{\beta_k \beta'_k}{1 - R^2_{k,t}}. \quad (23)$$

$$u = \frac{-\beta_k}{1 - R^2_{k,t}}. \quad (24)$$

$$d = \frac{1}{1 - R^2_{k,t}}. \quad (25)$$

From (12), (13), (23), (24), and (25), we have

$$p^{-1} = \frac{1}{1 - R^2_{k,t}} \begin{vmatrix} (1 - R^2_{k,t})r^{-1} + \beta_k \beta'_k & -\beta_k \\ -\beta'_k & 1 \end{vmatrix}. \quad (26)$$

From (2), (3), (6), and (26),

$$\beta = \frac{1}{1 - R^2_{k,t}} \begin{vmatrix} (1 - R^2_{k,t})r^{-1} + \beta_k \beta'_k & -\beta_k \\ -\beta'_k & 1 \end{vmatrix} \begin{vmatrix} r_c \\ r_{ck} \end{vmatrix},$$

or,

$$\beta = \frac{1}{1 - R^2_{k,t}} \begin{vmatrix} (1 - R^2_{k,t})\beta_c + \beta_k \beta'_k r_c - \beta_k r_{ck} \\ -\beta'_k r_c + r_{ck} \end{vmatrix}. \quad (27)$$

From (2), (5), and (27),

$$R^2_{c.T} = \frac{1}{1 - R^2_{k.t}} (r'_{c, r_{ck}}) \left\| \begin{array}{c} (1 - R^2_{k.t}) \beta_c + \beta_k \beta'_k r_c - \beta_k r_{ck} \\ - \beta'_k r_c + r_{ck} \end{array} \right\|,$$

or,

$$R^2_{c.T} = \frac{1}{1 - R^2_{k.t}} \times \left((1 - R^2_{k.t}) r'_c \beta_c + (r'_c \beta_k)^2 - r'_c \beta_k r_{ck} - \beta'_k r_c r_{ck} + r^2_{ck} \right). \quad (28)$$

Substituting from (8) in (28) and simplifying,

$$R^2_{c.T} = R^2_{c.t} + \frac{(r'_c \beta_k - r_{ck})^2}{1 - R^2_{k.t}}. \quad (29)$$

Solving (29) for r_{ck} , we have

$$r_{ck} = r'_c \beta_k \pm \sqrt{(R^2_{c.T} - R^2_{c.t})(1 - R^2_{k.t})}. \quad (30)$$

Formula (30) gives explicitly the value which r_{ck} must take to yield a specified increase in the multiple correlation coefficient. For computational convenience, the right-hand side may be written somewhat differently. The term $r'_c \beta_k$ implies the computation of the multiple regression coefficients for predicting test k from the test battery. This is not necessary if we have given predicted criterion scores from the test battery with test k excluded.

Let:

- z = the matrix of standard test scores, excluding test k ,
- K = the vector of standard scores for test k ,
- C = the vector of standard criterion scores,
- C_s = the vector of predicted criterion scores when test k is excluded from the battery, and
- r_{ks} = the correlation of test k with the predicted criterion scores.

By definition,

$$z \beta_c = C_s. \quad (31)$$

The correlation of test k with the predicted criterion score is

$$r_{ks} = \frac{\frac{1}{N} K' C_s}{\sqrt{\frac{1}{N} C'_s C_s}}. \quad (32)$$

Substituting (31) in (32),

$$r_{ks} = \frac{\frac{1}{N} K' z \beta_c}{\sqrt{\frac{1}{N} \beta_c z' z \beta_c}}. \quad (33)$$

By definition,

$$\frac{1}{N} K' z = r'_k. \quad (34)$$

$$\frac{1}{N} z' z = r. \quad (35)$$

Substituting (34) and (35) in (33),

$$r_{ks} = \frac{r'_k \beta_c}{\sqrt{\beta'_c r \beta_c}}. \quad (36)$$

Substituting from (6), (7), and (9) in (36),

$$r_{ks} = \frac{r'_c \beta_k}{R_{c,t}}. \quad (37)$$

Solving (37) for $r'_c \beta_k$,

$$r'_c \beta_k = R_{c,t} r_{ks}, \quad (38)$$

which may be substituted in (30).

Suppose now, we prefer to specify a desired increment in the multiple correlation rather than its square. If we indicate this increment by a , we have

$$R_{c,t} + a = R_{c,t}, \quad (39)$$

or

$$R^2_{c.t} - R^2_{c.t} = a(a + 2R_{c.t}). \quad (40)$$

Substituting from (38) and (40) in (39) gives

$$r_{ck} = r_{ks}R_{c.t} \pm \sqrt{a(a + 2R_{c.t})(1 - R^2_{k.t})}, \quad (41)$$

which is the same as Formula A.

Manuscript received 5/1/50

AN EMPIRICAL VERIFICATION OF THE WHERRY-GAYLORD ITERATIVE FACTOR ANALYSIS PROCEDURE*

ROBERT J. WHERRY

OHIO STATE UNIVERSITY

JOEL T. CAMPBELL

PERSONNEL RESEARCH SECTION, ADJUTANT GENERAL'S OFFICE

AND

ROBERT PERLOFF

OHIO STATE UNIVERSITY

A comparison of the Wherry-Gaylord iterative factor analysis procedure and the Thurstone multiple-group analysis of sub-tests shows that the two methods result in the same factors. The Wherry-Gaylord method has the advantage of giving factor loadings for items. The number of iterations needed can be reduced by doing a factor analysis of sub-tests, re-grouping sub-tests according to factors, and using each group as a starting point for iterations.

Wherry and Gaylord† proposed an iterative method of factor analysis which identified the factor structure of the test and gave factor loadings for each item, but which did not require item inter-correlations. This method has been empirically verified in factor analyzing an officer-qualification check list. Comparison of results from the Wherry-Gaylord with those from a Thurstone-type multiple-group analysis of sub-tests showed that the two methods give identical factors after rotation first to orthogonality and then for meaningfulness.

The analysis was based on ratings of each of 231 Regular Army officers by his immediate superiors. The check list used had 289

*This research was carried out under Contract No. WSW-2503, between the Department of the Army and Ohio State University. This paper is based on the final report PRS No. 827 under that contract. The opinions expressed herein regarding matters relating to the Department of the Army are those of the authors and are not necessarily official.

†Wherry, Robert J., and Gaylord, Richard H. The concept of test and item reliability in relation to factor patterns. *Psychometrika*, 1943, 8, 247-264.

items, each of which was marked on a five-point scale. For the analysis, unfavorable items were reflected, and each item was dichotomized as close as possible to the 50 per cent level of difficulty.

The original Wherry-Gaylord iterative analysis called for (1) computation of item-test coefficients, (2) grouping of items with highest coefficients into a new "test," (3) computation of coefficients between each item and the new "test," (4) addition of items whose coefficients increased and dropping of items whose coefficients decreased. These steps are repeated until stability is reached. Then those items which had been rejected are formed into a new "test," item-test coefficients are computed, and the procedure is repeated as many times as necessary.

Since the Officer Qualification Form had 289 items, it was felt that pre-sorting of items into "factor" piles might reduce the number of iterations if the number of factors turned out to be at all large. Accordingly, the items were sorted into 13 groups according to the following categories: ability, attitude toward work, efficient use of subordinates, force, general cultural level, knowledge of profession, military appearance, morality, originality, performance, relation to subordinates, relation to superiors, and sociability. There were from 8 to 48 items in each group.

Each of the 13 sub-tests was used in turn as a starting point for iterative analysis. The item-test coefficients computed were tetrachorics between marked-high—marked-low on an item and upper-half—lower-half in sub-test score. After approximately 4 iterations in each case, it was found that the items selected on the 13 scales fell roughly into three groups or patterns. Final sub-test scores for the 13 iterated scales were intercorrelated and each test was found to correlate at least .98 with every other test in one of the three groups or patterns.

Two further steps were taken to see whether other factors were obtainable. First, all items not appearing in the 13 final sub-test scales were used as a 14th scale. Iteration of this group quickly resulted in another scale duplicating Group III. Secondly, an examination was made of factor loadings and all items with loadings less than .30 on any of the three groups were selected as a 15th scale. Iteration of this group of items resulted in a sub-test which contained several new items. It was also tending to iterate toward one of the three groups, but iteration was stopped before it reached that

stage, and loadings in this test were used to represent a Group IV factor.

In order to obtain correlations among the four factors, items with high loadings on each scale were selected as a test of that factor. The subjects were scored on these sub-tests and correlations obtained between the sub-test scores. Using these correlations in lieu of actual inter-factor correlations, a transformation matrix was secured and applied to the item sub-test coefficients to obtain orthogonal factor loadings. About 20 of the 289 resulting communalities were above 1.00. Examination showed that in every case these were items used in defining the final sub-tests, resulting in upward contamination of their factor loadings. This contamination ($r_{\text{contamination}} = \sqrt{1/n}$) was removed, and the transformation matrix reapplied. This reduced all except 6 items to a communality of less than 1.00. For these remaining items a proportional decrement across the loadings was used to reduce the communalities to unity.

After orthogonality and communalities of unity or below were achieved, the four factors were rotated for meaningfulness.

The large number of items made complete plotting infeasible since the dots could not be labeled and seen. This necessitated selection of only the highest positive and negative combinations for plotting and the securing of actual rotated loadings mathematically. Each succeeding transformation matrix was superimposed on the preceding transformation matrix, and the resulting matrix applied to the original oblique loadings to secure the final orthogonal rotated loadings.

Descriptions of the four factors:

Factor I had its highest loadings on the following 10 items:

- .90 280. No attempt to help others (R)*
- .89 155. Lacking in sincerity (R)
- .82 207. Does not secure loyalty of subordinates (R)
- .80 171. Feels mistreated (R)
- .80 107. Hated by subordinates (R)
- .79 139. Harbors grudges (R)
- .78 130. Not cooperative (R)
- .78 153. Caustic in remarks (R)
- .78 204. Cannot apply knowledge (R)
- .78 237. Selfish in motives (R)

*"R" indicated item was reflected in scoring.

Reduced to a description, we have an officer who "is sincere, helpful, cooperative, and satisfied and who engenders liking and respect." This factor seems best described as *Proper Attitude Toward the Job*.

Factor II had its highest loadings on the following 10 items:

.92	42.	Establishes cordial relations
.91	100.	Well liked by fellow officers
.82	263.	Knows his subordinates
.78	19.	Affable and genial
.76	50.	Makes duty assignments according to ability
.75	53.	Gets along well with subordinates and superiors
.75	68.	Pleasing personality
.75	70.	Has vitality
.76	16.	Assigns men properly
.72	20.	Physical endurance

These items seem to describe an officer who "is genial, cordial, and well liked by subordinates, fellow officers, and superiors, and who handles his relationships with them in a satisfactory manner." This factor is therefore identified as *Successful Interpersonal Relationships*.

Factor III has its ten highest loadings on the following items:

.72	182.	Makes bold and quick decisions
.68	119.	Lacks ability (R)
.68	250.	Has little force (R)
.65	177.	Fails to exercise initiative (R)
.64	265.	Forceful
.63	93.	Good leader
.62	60.	Needs to assert himself (R)
.61	29.	Physically unimpressive (R)
-.60	225.	Quiet
.60	249.	Timid (R)

Boiled down to a thumbnail description, the items say such an officer "is bold, forceful, and quick to lead or take the initiative; never quiet, timid, or afraid to assert himself." This factor is therefore identified as *Forceful Leadership and Initiative*.

The ten highest loadings on Factor IV were

.79	12.	Does not know his job (R)
.76	260.	Competent
.75	241.	Persevering
.74	239.	Mentally alert
.73	201.	Makes little progress toward objectives (R)
.73	284.	Criticizes superiors in front of junior officers (R)

- .70 281. Not well informed concerning his duties (R)
- .69 165. Ignorant (R)
- .69 199. Shirks responsibility (R)
- .66 65. Seeks easiest assignments (R)

These items clearly depict an officer who "is competent, alert, informed, and persevering; one who gets things done and likes to do them." This factor is therefore identified as *Job Competence and Performance*.

Since it seemed that four factors were quite few to describe almost 300 items, and since the original thirteen sub-tests had "appeared" to be different, it seemed wise to run a factor analysis of the scores on the 13 selected original sub-tests.

Table 1 shows the intercorrelations among the 13 original sub-tests before iteration. A group factor analysis* yielded 4 factors. Rotated first to orthogonality and then for meaningfulness, the loadings are those shown in Table 2, with residuals in the upper half of Table 1.

Factor I had its highest loadings on

- .69 12. Relation to Superiors
- .58 2. Attitude toward work
- .54 8. Morality
- .39 1. Ability
- .35 10. Performance

This factor is quite easily identified as being the same as Factor I from the iteration process, namely *Proper Attitude Toward the Job*.

Factor II had its highest loadings on

- .61 11. Relation to Subordinates
- .60 3. Efficient use of Subordinates
- .50 7. Military appearance
- .57 13. Sociability
- .44 9. Originality
- .41 8. Morality

Again the factor is clearly identifiable as the Factor II from the iterative process, namely *Successful Interpersonal Relationships*.

Factor III had its highest loadings on

- .75 7. Military appearance
- .57 1. Ability

*Thurstone, L. L. *Multiple Factor Analysis*. Chicago: Univ. of Chicago Press, 1947.

.43	4.	Force
.39	6.	Knowledge of Profession
.38	9.	Originality
.38	5.	General Cultural Level
.38	3.	Efficient Use of Subordinates
.35	10.	Performance

Inspection of both the topics and items contained in Factor III from the iterative process (the 10 highest items came from sub-tests 1, 4, 7, 9, 10, 11, and 13) indicate the identity of this factor with that one. It is accordingly labelled *Forceful Leadership and Initiative*.

Factor IV had its highest loadings on

.77	10.	Performance
.75	6.	Knowledge of Profession
.74	9.	Originality
.67	11.	Relation to subordinates
.63	2.	Attitude toward work
.61	4.	Force
.57	1.	Ability

Again this factor is immediately seen to correspond to Factor IV on the iterative process namely, *Job Competence and Performance*.

This finding clearly points out that such initial groupings, factor analyzed, and regrouped according to factors, rather than by topics, would form better initial breakdowns for starting the iterative procedure, and lead to further time saving in that process. For example, in this case, the four factor divisions of sub-tests would have cut the iterative sequences from 13 to 4 and would have led more quickly and with equal accuracy to the same result.

Both the Wherry-Gaylord iterative analysis and the Thurstone group method, however, gave the same factors.

The effectiveness of the suggested use of sub-tests as a short cut will obviously be a function of the homogeneity of the sub-tests editorially selected. The consequences of varying degrees of homogeneity are indicated below:

(a) If completely heterogeneous (implying no success in editorial judgment), factor analysis would yield *one* factor and we would thus have the original Wherry-Gaylord suggestion of using *total test score*. In this instance, nothing would be gained by using

the suggested short cut, while the only loss would be in terms of time spent.

(b) If partly homogeneous with some heterogeneity (implying some editorial success) we would expect results similar to those in the present study.

(c) If completely homogeneous within a sub-test but heterogeneous with respect to certain other sub-tests (better editorial acumen than that represented by the present study) the factor analysis would yield our findings in clearer form (less high loadings on other factors for sub-tests). This would be nicer looking, but would make no practical difference.

(d) If each test was completely homogeneous both internally and with respect to the other subtests (implying perfect editorial ability), factor analysis would yield no common factors and would indicate that each sub-test is already an independent factor. (Note: If the same is true of a single sub-test, it will not have loadings on any of the common factors found, but would be included as a separate factor.)

Manuscript received 5/15/50

Revised manuscript received 8/10/50

TABLE 1
Intercorrelations and Residuals for 13 "Guessed" Factors*

Sub-Test	1	2	3	4	5	6	7	8	9	10	11	12	13
1 Ability	—	00	-02	-03	05	01	-06	03	00	00	01	-03	-01
2 Attitude	72	—	02	05	-04	-01	06	-03	01	01	01	-01	-03
3 Efficient Use of Subordinates	57	61	—	01	-05	04	-09	00	-03	07	09	02	00
4 Force	64	68	72	—	-06	-01	07	02	-01	00	00	04	-03
5 General Cultural Level	63	51	52	53	—	00	03	01	02	-05	-04	01	04
6 Knowledge of Profession	76	72	71	73	61	—	00	00	-01	-03	-02	-03	-02
7 Military Appearance	43	41	56	65	49	47	—	01	-07	08	00	03	-05
8 Morality	60	71	58	59	49	61	37	—	-01	-04	01	-02	04
9 Originality	71	75	75	79	66	81	52	63	—	03	-07	07	00
10 Performance	79	82	73	75	57	80	51	64	86	—	-01	-02	-06
11 Relation to Subordinates	57	71	83	70	52	69	47	71	74	73	—	04	-02
12 Relation to Superiors	64	81	56	60	51	63	33	73	69	73	71	—	-03
13 Sociability	42	58	59	51	47	51	33	67	82	51	70	59	—

*Intercorrelations are shown below the principal diagonal, and residuals above. The decimal point has been omitted in all entries.

TABLE 2

Factor Loadings*

Sub-Test	I Conscientious Attitude	II Personal Relations	III Force and Initiative	IV Job Performance	h^2
1	39	08	57	57	81
2	58	23	21	63	83
3	11	60	38	49	76
4	12	39	43	61	72
5	17	27	38	49	49
6	22	22	39	75	81
7	03	58	66	10	78
8	54	41	09	49	71
9	10	44	38	74	90
10	35	20	35	77	88
11	22	61	07	67	87
12	69	25	12	54	84
13	34	57	—01	45	64

*The decimal point has been omitted from all entries.

THE CENTRAL INTELLECTIVE FACTOR*

H. J. A. RIMOLDI

UNIVERSITY OF CHICAGO

The proof of the existence of "g" is more than a methodological problem and concerns the very core of psychological theory. The principles of noegenesis should be identified experimentally before a final opinion can be rendered about "g." Many general factors isolated in different studies are not necessarily "g." In the present study a second-order unrotated general factor has been identified by using Thurstone's method. It seems possible to identify this factor with "g." In the first order, factors that seem to represent the first and second principles of noegenesis have been found. The existence of synthetic and analytic activities and their interplay in intellectual performances is indicated. The relation of likeness is of great interest in explaining cognitive abilities and is isolated both as a first and second order factor. For the final identification of factors the search should be conducted beyond the elementary listing of tests. The dynamic aspects underlying factors are more meaningful than their simple description. The second order gives indications that allow for a better interpretation of fundamental psychological activities.

Introduction

Since Spearman defined "g" many factorial studies have been published in relation to intelligence. No satisfactory agreement has yet been reached on issues germane to the problem of the nature of intelligence, in spite of the different methods and tests tried. Some factorialists have often been satisfied with the simple enumeration of variables. Nevertheless, factor analysis has more to its credit than merely cataloguing factors.

Spearman's interest was to delimit and define "g" as a general factor best expressed by the principles of noegenesis† and by what

*The final part of this study was carried on at the Psychometric Laboratory, University of Chicago, during the year 1946-47. This part of the research was completed under a State Department Grant and a Frank Fund Fellowship.

The author is indebted to Dr. L. L. Thurstone for his assistance and to Mr. V. S. Tracht for his help in preparing the manuscript.

†As described by Spearman the principles of noegenesis refer to: 1) "a person has more or less power to observe what goes on in his own mind," 2) "when a person has any two or more ideas (using this word to embrace any items of mental content, whether perceived or thought of), he has more or less power to bring to mind any relations that essentially hold between them," and 3) "when a person has in mind any idea together with a relation, he has more or less power to bring up into mind the correlative idea" (27). The theoretical postulation and a complete discussion of these three principles is found in *The Nature of Intelligence and the Principles of Cognition* (25).

he calls abstraction, adding that the analytic procedures "tend to load noegenetic processes with 'g'" (30). His interest in other factors was secondary and he cautioned against those conditions that would disturb the tetrad criterion and make "g" disappear. Proof of the existence of "g" were given by Holzinger (17) and Brown (6), among others.

Some British psychologists and Spearman himself spoke of overlapping factors, and in this endeavor mainly the verbal and the space factors were accordingly defined (31, 32, 33, 13). "g" was interpreted as "general fund of energy," will power, maturation, neural plasticity, condition in the blood, neural energy, chance, and so on (16, 27, 30).

In Thurstone's theory the existence or non-existence of a general factor is not previously postulated. The main interest is to discover the number and properties of the factors that reproduce the given raw data, mainly the correlational matrix (38). As several experimental studies have shown, if there is a general factor it will become evident by using Thurstone's method (4, 5, 15, 36, 39).

The early procedure of rotating the factors while keeping them orthogonal was modified by introducing oblique factors. These are linearly independent but statistically related. The study of the second order—that is the analysis of the correlations between the primaries—has been little explored up to the present. The statistical and methodological implications have been partially discussed (24, 36, 38), and probably it is in the second order where the interaction of primary factors is to be further clarified. Nevertheless, any second order findings should be carefully interpreted on account of our lack of experience and of the theoretical and methodological assumptions involved.

Thurstone (35, 36) described several primary factors and his pupils and associates in a series of different studies confirmed their existence and characteristics. Some of these factors have been lately split into several others, or their properties redefined, for instance the perceptual and the space factors, etc. (37).

It is customary to present the results obtained by Thurstone *et al.* as incompatible with those reached by Spearman and associates. Since our problem is related to this, we shall review some of the pertinent bibliography on the subject.

Blakey (4), using Thurstone's method, reworked Brown and Stephenson's data and found a verbal, a space, and a perceptual speed factor, plus another variable that may be interpreted as Spearman's

central intellectual factor or as the effect of maturation on the subjects. In a study of non-verbal tests the same author (5) reports a factor which is general for the special battery employed.

A general factor and others were also found by Wright (39) in analyzing the Stanford Binet scale. This author believes that this factor is due to the effects of maturation rather than "*g*." This interpretation was criticized by Burt and John (10) who, by analyzing the Terman Binet scale, discovered a general intelligence factor which may "be regarded as indicating the particular characteristics which the tests were designed to measure." Whether this interpretation is accepted or not, it is related to Burt's theoretical position, wherein factors are mainly principles of classification (9, 11).

Thurstone (36) reported that the correlations among the primary factors indicate that "each of the primary factors can be regarded as a composite of an independent primary factor and a general factor which it shares with other primaries." Swineford (34) stated that this second-order general factor is similar to Holzinger's general factor.

Carroll (12) described a second-order general factor common to all the primaries, and Goodman (14) reported a similar result. Mellone (20) also described a general factor—present in 7 years old boys and girls—by using Thurstone's method and orthogonal rotations, and Balinsky (2) found a general factor, which the author calls "*g*" at 9 and 50-59 years of age.

Alexander (1) reports several factors, one of them common to all the tests in the battery. But Alexander's criterion for rotation was to pass an axis through a cluster of tests previously interpreted as tests of "*g*." The lack of an independent criterion for the determination of "*g*" makes his conclusions less valid than otherwise. When Yela (40) analyzed Alexander's correlational matrix by means of Thurstone's method, he did not find a general factor in the first order. The second order indicates the existence of a general factor common to the cognitive abilities described in the first order.

It is interesting to report that Spearman (28) in reworking Thurstone's data, found a general factor plus a Verbal (*V*), a Space (*S*), a Number (*N*), and a Memory (*M*) factor. Induction (*I*), Reasoning (*R*), and Deduction (*D*) did not appear in the results.

Reyburn and Taylor (22) advanced the idea that "*g*" might not represent a single factor.

A summary of these paragraphs indicates:

(1) There is no basic disagreement in relation to factors such as *V*, *M*, *S*, etc.

(2) A general factor has been found by using Thurstone's method of analysis.

(3) This factor does not always agree with the definition of "*g*" as abstraction and noegenesis.

(4) The characteristics of "*g*" as given by Spearman and associates, the results of experimental work on the Primary Mental Abilities—mainly factors *I*, *R*, and *D*—and the findings of Spearman (28) and others seem to indicate that, experimentally considered, "*g*" might not be a unitary factor and that from the theoretical standpoint the study of the relationships between some of the Primary Mental Abilities and "*g*" may prove valuable in explaining the dynamics of intellectual performances.

In relation to this last point we planned the present study. The study of the second order might give valuable clues for the interpretation.

It is difficult to find a large battery of tests that will fulfill the tetrad criterion. The conditions required have been stated by Spearman and his associates. The general factor found by using techniques other than Spearman's should not necessarily be equated to "*g*." The proof or disproof of "*g*" is more than a mathematical matter, as indicated by Brown and Stephenson (7). Consequently it seems that its acceptance or rejection pertains to psychological theory. The crux of the problem is: Are the laws of noegenesis and abstraction, as understood by Spearman, proved facts or do they still belong to the realm of theory? If these concepts are realities in an empirical sense they should be experimentally verifiable, whether we use factor analysis or any other convenient psychological technique.

The battery of tests employed for such a study should have tests of "*g*," *I*, *R*, and *D*, avoiding insofar as possible the presence of variables that may mar the results on account of their complexity. We thus selected tests that were considered by previous authors as good measures of these factors.

The Tests

Tests No. 1 and No. 2: Geometrical Forms No. 1 and Geometrical Forms No. 2, respectively, similar to the Otis and Thurstone's tests (36) of the same name. Spearman reported that these tests are saturated in "*g*" and Thurstone has shown that they are loaded in the *I* and *S* factors. The difference between these two tests lies in the fact

that in the first one there is a larger use of words than in the second one.

Test No. 3: Number Series. Thurstone (35) reported the high saturation of this test in the *I* factor and in *N*, *V*, *D*, and *M*.

Test No. 4: Verbal Analogies. Spearman reported the saturation of this test in "*g*" and Thurstone indicated high loadings in *V*, *P*, and to a less extent in *S*.

Test No. 5: Pedigrees. This test is similar to the one used by Thurstone (36) who reported its loadings in *I* and in *M*.

Test No. 6: Inventive Synonyms. This test has been considered as a good measure of "*g*" and Thurstone (35, 36) indicated that it is loaded in *V*, *W*, *P*, and *D*.

Test No. 7: Group of Figures. Taken from the "City and County of Newcastle-upon-Tyne Education Committee Intelligence Test." This test is similar to Thurstone's Figure Grouping test (36) reported by this author as loaded in *P* and *I*. Several rows of drawings are given and the subject has to mark in each row the design that is different from the rest of the designs in the same row.

Test No. 8: Classification of Figures. Taken from the same battery as the previous test. This test consists of several rows of designs divided into three parts. The designs in the first and second parts are equal in some way. In the third part four designs are given, two of them similar to the drawings in the first part of the row and two similar to the drawings in the second part. The subject has to give a number—1 or 2—to each of the four designs according to their similarity to the designs in the first or second part of the row. Thurstone (35) reported this test as loaded in *V*, *R*, and *W*.

Test No. 9: Arithmetical Reasoning. Similar to Thurstone's (36) and to Burt's (8) tests of the same name. Spearman indicated that tests of this kind are essentially loaded with the "relation of conjunction" and are given as "*g*" tests. According to Thurstone (36) it is loaded in *R*, *I*, *V*, and *N*.

Test No. 10: Absurdities. Similar to Thurstone's test of the same name (36) with loadings in *V* and *I*.

Test No. 11: Group of Letters. Thurstone (36) has shown its saturation in *I*.

Test No. 12: Code. This test was taken from the "City and County of Newcastle-upon-Tyne Educational Committee Intelligence test." A number of designs are given, each one indicating a particular letter.

In the lower parts these same designs are repeated, although different parts have been taken from them. The subject has to indicate to which letter each of the designs corresponds. Tests of this nature have been regarded as correlated with "g."

Test No. 13: Secret Writing. Similar to Thurstone's test of the same name. (36). This variable shows loadings in *I* and to a lesser extent in *S* and *M*.

Test No. 14: High Numbers. Similar to Thurstone's test of the same name (36). This author has indicated its saturation in *N*, *S*, and *I*.

Test No. 15: Numerical Judgment. Taken from Thurstone's Primary Mental Abilities (35), where it shows loadings in *R*, *I*, and *N*.

Test No. 16: Three Higher. Taken from Thurstone's Factorial Studies of Intelligence (36) where it is mainly loaded in *N* and *I*.

Test No. 17: Letter Series. Thurstone has shown its saturation in factor *I* (36) and it also has been reported as a good measure of "g."

Test No. 18: Directions. Similar to Thurstone's test of the same name (36) which is loaded in *I* and *V*.

Test No. 19: Areas. Similar to Thurstone's test of the same name (35) which is reported as saturated in *I*, and to a lesser extent in *R*, factor 11, *M*, and *S*.

Test No. 20: Form Analogies or Pattern Analogies. This test was reported as loaded in "g" and Thurstone indicated (35) its saturation in *P*, *I*, *R*, and *D*.

Test No. 21: Number Pattern. Taken from Thurstone's Factorial Studies of Intelligence (36) where it is given as a test of *I* and *P*.

Test No. 22: Inventive Opposites. It has been considered as loaded in "g." Thurstone indicated its saturation in *V*, *W*, *R*, *M*, and factor 10.

Test No. 23: Reasoning. Taken from Burt (8). This test is similar to other tests of the same name, where problems of different complexity are verbally presented.

Test No. 24: Reasoning and Inferences No. 1. This test is similar to the Reasoning test (Syllogisms) as given by Thurstone (35). This author has shown its loading in *D*, *V*, *R*, and *I*.

Test No. 25: Reasoning and Inferences No. 2. Similar to Thurstone's test of reasoning (36). According to this author is loaded in *N*.

Since many of our tests are not pure tests—in a factorial sense—other factors than *I*, *R*, and *D* and "g" are present, as can be corrob-

orated by studying the corresponding tables given by Thurstone (35, 36).

Since our population consisted of Spanish speaking persons, some of the tests were translated into Spanish, when a translation was not already available. Before administering the test to the population here studied, the whole battery was tried with a smaller group of 50 children, of both sexes and between 11 and 14 years of age, to determine the more convenient norms of administration.

Population

Three hundred eighty-four children between 11 and 14 years of age were administered the whole battery in four different sessions, each of a duration not longer than 45 minutes. The whole testing was completed for each group of approximately 40 subjects within a week. Half of our subjects were boys and half girls, each age and sex being represented by the same number of subjects.

The selection of the population was done as follows. From the complete school population of a city of approximately 300,000 persons, 12 different schools were randomly selected. From each of these 12 schools a complete list of the 11 to 14 year-old students in the three more advanced grades was secured, and from these lists our group of 384 subjects was selected at random. There was an equal number of subjects for each age level, for each grade, and for each sex. The testing was performed by two persons during the year 1943. A detailed description of the population, including the tests and the educational, age, and sex differences, together with the study of the test items and other norms has already been published (23).

*Method**

The correlations among the variables were analyzed by both Spearman's and Thurstone's methods. The results are given in Tables 2 and 3. To stabilize the communalities three successive factorizations were performed by using first the centroid method and afterwards, with the improved communalities, the multiple-group method twice in succession, each time using the preceding communality estimates. The final matrix of transformation and the oblique factor matrix are given in Tables 4 and 5, respectively. The correlations between the primaries, the orthogonal matrix for the second order,

*The statistical work including the first factorization was done with the help of the assistants of the "Instituto de Psicologia Experimental," University of Cuyo, Argentine, 1945-46. The author wishes to express his thanks to them for their valuable help.

the final matrix of transformation for the second order, and the correlations between the second-order factors are given in Tables 6, 7, 8, 9, and 10 respectively.

The Factors

Factor A:

Verbal Analogies	.34
Absurdities	.30
Code	.32
Three Higher	.28
Pattern Analogies	.25
Reasoning and Inf. No. 1.	.63
Number Pattern	.20
Secret Writing	.20

Analogies, Absurdities, and Reasoning and Inferences No. 1 involve problems that are verbally presented; Code and Pattern Analogies require the use of words to a lesser extent—the problems are presented graphically—while Three Higher requires the manipulation of numbers. This indicates that the nature of this factor goes beyond the mere presentation of the problems.

In Analogies a certain statement is given: "black is to white" and another similar incomplete statement such as: "red is to . . . blue, green, yellow, gray." The activity required implies a knowledge of the first condition and of the restrictions imposed upon the solution by the partially stated second condition. In all cases the subject has to combine and discern from among the different answers the one that completes each test item in the best way. The critical qualities of the terms to be related have to be isolated, but above crude analysis a further combination has to be performed among the conflicting statements.

In Absurdities various relations are given, the understanding of each one of them and of the whole system being a basic condition for a good solution. The correctness of the answer depends on both the analytical and synthetic activities and in the interaction of several different *Gestalts*.

The explanation given for Verbal Analogies is also valid in the case of Pattern Analogies. The problems are essentially of the same nature and the differences are due to the verbal presentation in the former. This last aspect is accounted for by the loading of Verbal Analogies in Factor F.

Reasoning and Inferences No. 1 is the most heavily loaded test.

To judge whether a conclusion is right or wrong it is necessary to analyze the premises and their interrelationships. Nevertheless the piecemeal study of each part is insufficient for the solution that requires a combination of the parts into a more inclusive whole. The same activity that was required for the solution of the previous tests is important here. Unless the subject can apprehend each particular sentence and the conflicting or non-conflicting statements in the whole syllogism, the correct answer is practically impossible. Again, the greater the flexibility of the subject in combining the parts in different ways, the easier the solution.

In Code the testee has to indicate to which of the complete patterns the partially given designs belong. He has to find the proper solution, due account being given to the distortions imposed upon the drawings; that is, one possible answer is discarded in terms of a better one.

In Three Higher the manipulation of numbers is required. One relation is given as the basis for the solution of the test; that is, any number three units higher than the previous one. The subject has to isolate from the consecutive numbers those that fulfill the stated condition; that is, some answers are discarded in favor of others, and so on.

Secret Writing and Number Pattern have a small loading in Factor A, but the two tests will be analyzed more in detail when examining Factor C.

From the previous description it seems that the processes underlying Factor A are not only structurally complex but also imply a complicated dynamic system. Not only analysis or synthesis are present but their interplay seems to be a fundamental characteristic in defining the properties of this variable.

Factor A seems to agree with what has been usually called reasoning. In our case it stresses the essentially dynamic character of the process, and mainly the plasticity required to perform such an activity in the rather complex situations here presented. In all these cases there are conflicting forces at play. The solutions will be easier for the subject the more plastic and consequently the more "*Gestalt free*" he is (24).

Factor C:

Secret Writing	.55
Geometrical Forms No. 2	.44
Number Patterns	.38
Geometrical Forms No. 1	.36

These four tests are presented in different contexts—words, numbers, drawings—indicating that the activity represented by Factor C transcends the immediate test material.

Overlapping figures are shown in both Geometrical Forms No. 1 and No. 2, and the subject has to indicate some of the possible relationships among these drawings; for instance, which is the number inside the rectangle but outside the triangle, and so on. In Geometrical Forms No. 2 the problems are stated with a minimum use of words. In a general instruction for all the problems the subject is told to mark the space surrounded by a solid line but outside the dotted lines. In both tests the common characteristic task is to locate exactly the point in which the figures relate to each other according to the given instructions, the items being previously defined.

In Number Patterns a series of numbers is given and for the solution of each item the testee must discover the relationship among the numbers. Here as in the two previous tests the key for the solution seems to be centered upon discovering or noticing the relationships among the parts. The latter is implicitly given in the problems although it is not explicitly stated.

In Secret Writing a similar activity is required. The testee has to discover in what manner certain letters are related to certain numbers. The problems here are more complex than in the previous tests. This variable is also loaded in Factor A.

In all these tests the subject must find the relationship among the parts. This activity is seemingly different from the one required in those cases where a relationship and one of the items are given. For instance, it is a different task to discover the relation between 8 and 9—8 smaller than 9—than to produce a number "greater than" 8. In the first case both items 8 and 9 are given and the person has to find how they are related. In the second case, one item—8—and a relation—greater than—are given and the subject has to find a solution that fulfills the condition of "greater than." It seems that the tests included in this factor are better represented by the former kind of activity. A careful testing of the hypothesis would not be a difficult problem. Tests of an exceedingly simple nature should be employed together with others of increasing difficulty, always keeping the crucial functions as free from contamination as possible.

Factor E:

Pedigrees	.87
Directions	.36
Reasoning	.20
Analogies	.20

Strictly speaking this factor appears in the present study as a doublet. Pedigrees and Directions are loaded only in Factor E, while Analogies and Reasoning show saturation in Factor F as well.

In Pedigrees a genealogical tree is presented and the subject has to answer problems such as: "John is married to" The subject is given an item—John—a relation—is married to—and has to discover the other item that will satisfy the relationship. In the tests representing Factor C the nature of the problem was different. In that case the subject had to discover the relationship between two items; here the relationship is clearly stated and its thorough understanding makes it possible to find the correct solution.

In the test of Directions the psychological activity involved is of a more complex although similar nature. Nevertheless, by the manner in which the statements are presented and by the nature of the problems, it is not as evident as in the previous case.

Our findings seem to indicate that it is possible to isolate experimentally the second from the third principle of noegenesis. Nevertheless it is difficult to prove or disprove this particular hypothesis until other tests than the ones here used are systematically applied.

The fact that Analogies is slightly loaded in this factor is suggestive since this test has been considered as a variety of "eduction of correlates." It is also suggestive that the reasoning test shows some saturation in Factor E.

Factor G:

Group of Letters	.41
Classification of Figures	.41
Numerical Judgment	.38
Group of Figures	.34
Letter Series	.26
Number Series	.25

In Group of Letters the subject has to indicate which group differs from the remaining ones. Some of the items—the easy ones—can be solved in a rather intuitive way. Probably some kind of analytical activity is involved in the solution; but, at least in the easy items, the psychological performance does not seem to require a great deal of analysis. This is not so in the last and more difficult items of the test. The subject seems to reach a solution by using the activ-

ities that have been described by Mieli (19) as "complexité" and "globalization."

We could apply a similar reasoning to explain the loadings in both Letter Series and Number Series. In the easy items the analytical activity is limited and the grouping of the letters or numbers is rather obvious. As soon as the problems increase in difficulty the grouping becomes less obvious, and the subject has to discover the structure characterizing each item.

Classification of Figures requires the grouping of similar designs according to a previously given norm. The solution is obvious in most of the items and the structures that the subject has to distinguish are of a simple nature. Few items require, if at all, a complex analysis.

In Numerical Judgment, among the several possible answers to an arithmetic problem, the testee has to select the correct one without wasting time working in detail the exact solution. The subject, according to the instructions, has to realize that an operation like 3.90 by 2.10 gives a result near to the multiplication 4.00 by 2.00 . Among the different given answers he has to select the one nearer to 8.00 . As a matter of fact, if the testee makes a detailed and careful study of each problem, this procedure would probably imply a definite fall in the score.

In Group of Figures, which is also loaded in Factor D, the testee is presented with several rows of figures. In each row one of the figures is different from the rest and the subject has to underline it. The element that disturbs the whole is very obvious and the solution, in practically all the items but the last few, becomes evident without any careful analysis of the elements.

On account of this it seems possible to suggest that this factor indicates the ability of bringing the parts together into a meaningful solution. Thus this variable agrees with Mieli's "globalization." When the problems become more difficult, a greater amount of analysis may be necessary. Nevertheless in our case and by studying the items that have been solved by our subjects, it is quite likely to assume that no complex activity was involved to any great extent.

It is interesting to observe that our findings agree with Mieli's opinion in the sense that "globalization" and "complexité" may go together, "complexité" implying the activity involved in the solution of more complex structures.

Factor F:

Arithmetic Reasoning	.45
Opposites	.39
Synonyms	.33
Absurdities	.33
Syllogisms	.34
Geometrical Forms No. 1	.46
Reasoning	.24
Analogies	.21

All the tests defining Factor F require the use of words. Analogies is also loaded in Factors A and E. Absurdities has saturation in factor A as well. Synonyms and Opposites are also loaded in Factor D, while Reasoning is saturated in E.

It is interesting to observe that, while Geometrical Forms No. 1 and No. 2 cluster together in Factor C, only the first one of these tests is loaded in F. The main difference between these two tests seems to be in the greater use of verbal expressions in Geometrical Forms No. 1.

Arithmetic Reasoning has been previously described by Thurstone as loaded in the Verbal factor. It is interesting to observe that the two tests involving arithmetical computations do not cluster together in any of our factors. It seems in view of this and of previous evidence that arithmetic ability may depend on functions different from the mere manipulation of numbers.

It is curious to find that neither Reasoning and Inferences No. 1 or Directions have loadings in Factor F. The interpretation of this fact as well as the interpretation of factorial findings in general should be done in terms of the population of tests and of individuals. A factor may be absent in a factorial study, but this does not indicate that it is not influential in the solution of the test. It is obvious that the basic hypothesis and the variables employed to test a fact are of fundamental importance in the final results and interpretations.

Factor D:

Synonyms	.68
Group of Figures	.59
Opposites	.32
Letter Series	.27

This factor transcends the immediate test context. Synonyms and Opposites are loaded in the verbal factor F, while Group of Figures and Letter Series show saturation in Factor G.

In Synonyms and Opposites the subject has to find the answer which stands in a particular relation to the given words. This relationship is that of likeness and its opposite.

The same reasoning can be applied in Group of Figures, where the subjects select the design that is different from the remaining ones. In Letter Series the subject forms a group similar to the given ones.

The relation of likeness and its opposite has been considered by Spearman as furnishing "the main resource of all mental tests, whether sensory or otherwise." Line (18) has stressed the importance of this relation; and in a previous study (24) we discussed the part that this relation plays in intellectual performances. As indicated there, the relation of likeness makes possible the extension of conceptual thinking to levels of high complexity. Applied to fundamentals of a concrete or abstract nature it allows for an endless succession of further developments.

The conclusion that two or more things are equal or different may be reached either in a purely perceptual level—red different from green—or in levels of high complexity, as when the equality of two scientific concepts is discovered. Both the "as" and "that" types of cognition are permeated by this special and somewhat unique relationship.

Factor B:

High Numbers	.52
Areas	.51
Number Patterns	.26

It is difficult to define this factor. In High Numbers the subject has to indicate which are those numbers higher than those placed at both sides. In Areas the testee has to add several white squares drawn on a larger black square, and state which is the total number of white squares in each item. Number pattern has already been described.

All these tests require the use of numbers but it is impossible to individualize this factor with certainty. A perceptual element may be present in the solution of these tests, mainly in High Numbers and Areas, or perhaps the existence of a space component could also be defended. In any of these cases however we should expect a different distribution of the loadings of the other tests.

The Second Order

The factorization of the correlations between the primaries in-

dicates that these intercorrelations can be explained in terms of three linearly independent factors (Table 7). The loadings of the tests in the second-order orthogonal factors are given in Table 11.

By plotting the loadings of the tests in the first second-order factor against "*g*" loadings we find a very close agreement as indicated in Figure 1. The differences found may be attributed, in many of these values, to rounding errors.

"*g*" loadings were obtained by using "*g*" reference tests and by means of Thurstone's formula (38). The differences between these two procedures were insignificant and the latter values are employed here.

By plotting on the sphere the orthogonal system obtained by factorizing the correlations between the primaries, the rotated second order—Table 9—was obtained. The matrix of transformation and the correlations between the second-order factors are indicated in Tables 8 and 10.

Factor α is orthogonal to the rest of the second order factors. Factor D is, of all the primary factors, the one more heavily loaded in α . Factors A, C, and E also show a small saturation. It is interesting to discover that factor D (likeness) is the one characterizing this second-order factor, and it is also understandable how the relation of likeness may permeate the activities implied in A, C, and E.

Factor β is positively correlated with γ . Factors G, C, and E are loaded in this factor. C was described as the capacity of finding relations, while E was tentatively interpreted as eduction of correlates. These two activities seldom occur in isolation, at least in ordinary test material; but on the contrary the solution of the test problems requires the interplay of both activities. Factor α would refer to a particular kind of relation, that of likeness, while factor β would refer to the mechanisms involved in dealing with relations and eduction of correlates.

In Factor γ cluster Factors A, B, F, and G. F was interpreted as a verbal component, B was not interpreted, A was characterized as an indicator of the plasticity implied in reasoning processes, and G apparently indicates the ability of bringing the different parts together into a meaningful solution.

Although factors A and G may be considered of a similar kind, they are nevertheless sufficiently different. One of them indicates plasticity—that is trying different kinds of combinations when some restriction is involved—while the other seems to refer mainly to the process of synthesizing the parts into a solution.

On account of this description it appears that Factor γ is related to the synthetic process, including here all the possible reshaping and redistribution of the elements and the use of the "instruments" by means of which this process is accomplished. For instance it is suggestive to find that the verbal factor has only loadings in γ as suggesting that the activity defined by the latter develops in a concrete way and by using whatever symbols the subject may have at his disposal.

Discussion of Results

It is known that factorial results should be interpreted in terms of the special test configuration. In our particular case the tests were selected with a certain purpose and thus it is not strange that we do not find factors like P , S , and so on in our final results. In a way we eliminated them deliberately from our battery. It should also be noted that this process of elimination can not be perfectly accomplished, pure tests being still an ideal aim in factorial studies.

The presence of a verbal factor is easy to understand and was expected in view of our variables. Induction — I — has been defined by Thurstone (35) as the ability to "find a rule or principle for each item of the test" and R —reasoning—as "successful completion of a task that involves some form of restriction in the solution" while D is of "deductive kind." Moreover the same author says that "the tests which Spearman has designed as best measures of his general factor seem to be inductive in character."

The fact that there is a general orthogonal second order factor where the tests show the same loadings as in " g " is an interesting finding. Its identification with " g " is only possible when the tests involve noegenesis and abstraction, which is the case in the present study.

This general factor does not explain all our values, since in the second order three factors are necessary to explain the corresponding intercorrelations, and since some of the residuals of the original correlational matrix after partialling out " g " are still appreciably high. Moreover, it is necessary to remember that the correspondence of the " g " loadings is for the unrotated second order, as should be expected in terms of Spearman's method.

Taking into account the fact that the tests employed are loaded in Thurstone's I , R , and D , we have some grounds to believe that these factors are related to " g ." Studying the characteristics of the tests employed and the definitions of the factors mentioned above, this

findi
Spea
bein
show
ders

poss
Reyh
" g "
have
pone
of th

quite
agre
lated
toldt
cussi
ted i
lation
in th

terpr
which

bring
subje
ably
influe

"glob
lation
the pr
than t
we ar

T
stance
is spli
import
F
ness.

finding is not surprising. Thurstone (35), indicates that most of Spearman's tests of "*g*" seem to be inductive in character—induction being understood in terms of his *I* factor—while Spearman (28) has shown how the functions involved in tests of this nature can be understood in terms of "*g*."

If these results are accepted, the second problem is related to the possibility of considering "*g*" as non-unitary, in the sense in which Reyburn and Taylor (21) referred to it. These authors believe that "*g*" can be split into several independent factors. In our case we have found seven factors in the first order, one of them a verbal component, and another one, Factor B, not precisely identified, neither of them understandable in terms of "*g*."

Some of the factors isolated in this study seem to correspond quite closely to variables identified by others. Factor A seems to agree with Mieli's "plasticité" (19) and with a similar variable isolated by Thurstone (37) as Factor *E*, and by ourselves (24). Bechtoldt (3) described a Factor *Y* similar to the one we are here discussing and Yela (40) found a factor called *Z*, that can be interpreted in the same terms. The fact that Yela's factor has a high correlation with his *R* factor—related to the analytical procedure followed in the solution of a test—deserves consideration.

Factor A appears in the second order loaded in factor α —interpreted as likeness and its opposite—, and to a greater extent in γ , which seems to indicate essentially a synthetic process.

It seems that Factor A should be interpreted as the capacity of bringing together several conflicting *Gestalts*, "the more plastic the subject, the more likely he is to be Gestalt free" (24), and it is probably related to some temperamental and personality concepts. The influence of this factor in reasoning may be of a fundamental kind.

Factor G was interpreted as similar to Mieli's "complexité" and "globalization" (19) and to the factor defined as "perception of relations in space necessary for the construction of a whole" (24). In the present case this definition should be extended to contexts other than the spatial one. Thurstone's Factor *A* (37) is similar to the one we are here describing.

The solution of an intellectual task requires, in particular instances, a greater amount of analysis, which explains why Factor G is split into Factors β and γ of the second order. Nevertheless the most important activity here involved seems to be that of synthesis.

Factor D has been described as perceiving the relation of likeness. We have dealt with this problem in a different study (24). It

is interesting to discover that this particular kind of relation appears as defining one of the second order factors, mainly α . If this is corroborated in further studies, it may give interesting clues as to the nature of intelligence and of the processes of reasoning at different levels of complexity. Probably the dynamic interplay of this factor with others could be followed and clarified by a systematic study of pathological cases. The basic research concerning this relation of likeness still requires a great deal of preparatory work.

Factors C and E appear, together with Factor G, in Factor β . The first one seems to indicate the ability of finding relations, that is what Spearman considers the second noegenetic principle. This explains in part how it is saturated in Factor α .

Factor E was tentatively interpreted as the ability to educe correlates and is also loaded in Factor β , with some small saturations in γ and α . While the variables in Factor γ seem to deal with more concrete psychological performances, those in β seem to refer mainly to the logical process underlying intellectual activity.

Factor β is positively correlated with γ . This may explain the loading of G in both factors.

The findings that Factors A, F, and G cluster in Factor γ is interesting since it seems to indicate that this represents mainly a synthetic activity. In this respect γ seems to represent the centuries-old concept of synthetic sense as enunciated first by Aristotle, developed in the scholastic psychology, and reintroduced in a somewhat modified form by the Gestalt school and by Moore and Moynihan (21) in later times. These last authors conclude that many tests involve the synthetic functions mainly when there is a perception of relations.

The fact that the verbal factor appears in γ may indicate the function that our means of expression play in intellectual activities.

If the second order refers to a more fundamental level of psychological dynamics, our results seem to indicate that one relation—that of likeness—is a basic step in the solution of intellectual problems; that analysis and synthesis are the two main procedures at our disposal, one indicating the more abstract logical performance that is mainly involved in cognition of the “that” type, the other pointing to the combination of the different elements and their expression by means of appropriate symbols.

It is possible that individuals differ in the amount of these last two abilities and that their use depends, to a great extent upon training and other conditions. In this sense it is possible to conceive the feasibility of developing certain abilities. The process would essen-

tial
bols

for
to t
cent
fun
seen
and

tual
ticut
chol
thos
scen

of in
tion,
thes
field
thos
quite
thos
and
tal f
activ

*I
This i
possibi
range
and th

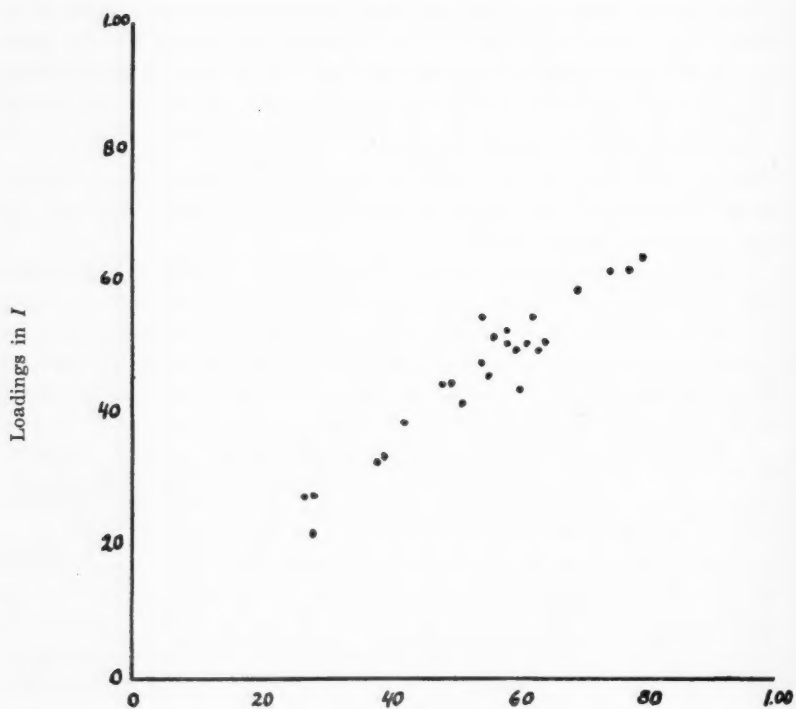
tially involve making more readily available the use of certain symbols. Cognition in this case evolves towards the so-called "as" type.

It seems also possible, if our findings are correct, to assume that for practical and theoretical purposes it would be more convenient to think of concepts such as "g" as non-unitary in character. The central intellectual factor should be considered in terms of more fundamental and dynamically interacting activities. These latter seem to agree with the results of previous studies of an experimental and non-experimental nature.

Some of the factors usually described as belonging to intellectual activity could well be only the results of test construction or particular training, and it is our opinion that in the description of psychological parameters a careful distinction should be made between those results which are due to the test context and those which transcend the immediate context of the tests.

Intellectual activity is a dynamic performance, and it would be of interest to know the effects of education, training, and maturation,* among others, upon the factors here described. The fact that these findings seem to agree with observations coming from other fields and that some of the factors here described seem to agree with those found by other authors and ourselves in different studies, is quite suggestive. Moreover, the identification of factors similar to those isolated in other countries with different cultural backgrounds and traditions is interesting in terms of the influence of environmental factors in factorial invariance, and in general in psychological activities.

*It has been suggested that maturation could explain part of our findings. This is a good hypothesis to test, but in this study no investigation of such a possibility was made for several reasons: among others, the fact that the age range here included is not large enough for a valid test of such a possibility and the fact that the study was not planned for such a purpose.



"g" Loadings

FIGURE 1

Plot of "g" Loadings against Loadings in the First Second-Order Factor.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25

"T

T

TABLE 1
Product Moment Correlations between the Tests*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1																									
2	30																								
3	35	21																							
4	35	22	45																						
5	36	18	31	50																					
6	32	08	35	56	12																				
7	15	08	31	22	26	57																			
8	30	10	32	34	27	22	13																		
9	64	15	55	57	37	44	29	34																	
10	32	18	31	48	39	47	13	26	42																
11	28	23	35	45	24	31	26	42	41	34															
12	31	12	40	48	32	36	22	28	42	42	35														
13	24	29	37	28	27	10	25	21	35	25	30	24													
14	20	18	26	32	19	15	12	25	30	14	21	30	18												
15	23	19	40	40	29	08	20	47	43	39	40	37	31	26											
16	27	10	46	62	32	43	19	25	53	38	35	42	25	37	46										
17	40	25	51	56	39	45	32	33	58	39	45	46	38	33	39	41									
18	33	17	34	40	69	40	14	25	45	31	27	33	32	22	35	37	40								
19	10	06	16	21	18	09	02	22	16	15	15	16	04	31	13	17	19	14							
20	32	15	36	45	30	38	21	24	40	30	37	43	26	25	32	35	42	28	14						
21	24	07	33	36	22	30	20	23	45	26	22	36	65	25	24	33	39	21	18	26					
22	29	09	24	55	16	57	27	35	59	44	25	42	21	30	38	45	49	39	18	29	28				
23	37	19	49	52	42	42	24	29	52	44	41	39	28	25	40	40	44	44	23	41	31	50			
24	04	02	21	29	03	24	07	08	20	26	17	20	14	06	25	17	23	17	05	22	17	18	18		
25	31	08	18	34	29	21	19	18	30	29	18	19	18	11	22	23	27	19	18	12	22	25	32	15	

*The decimal point has been omitted for all entries.

TABLE 2
"g" Loadings

Test	"g" Loading	Test	"g" Loading	Test	"g" Loading
1	.54	10	.60	19	.27
2	.28	11	.57	20	.56
3	.62	12	.61	21	.51
4	.78	13	.49	22	.63
5	.54	14	.42	23	.69
6	.60	15	.58	24	.28
7	.38	16	.64	25	.39
8	.48	17	.74		
9	.80	18	.58		

TABLE 3
The Centroid Factorial Matrix*

Tests	I	II	III	IV	V	VI	VII	h^2
1	35	19	50	18	10	-26	11	53
2	21	09	46	-11	-08	07	-08	29
3	58	12	22	14	-01	06	-11	43
4	68	14	08	28	24	09	09	64
5	37	20	25	06	68	21	-50	99
6	41	04	05	79	08	13	16	84
7	28	04	13	51	-20	16	-23	47
8	47	30	03	03	-09	-22	-13	39
9	61	15	32	35	16	-15	14	69
10	56	00	12	16	26	-13	10	45
11	56	17	13	-04	-08	-13	-21	43
12	60	10	04	16	09	11	03	42
13	42	01	56	-08	00	24	-08	56
14	33	51	05	02	-08	18	09	42
15	60	20	04	-03	00	-16	-16	45
16	58	16	00	24	13	06	10	45
17	62	18	24	28	-03	08	-04	56
18	48	10	21	13	28	11	-18	42
19	19	53	-09	-04	14	09	09	36
20	53	07	12	14	00	09	-02	33
21	40	20	33	09	-03	23	22	42
22	51	24	01	46	03	-13	23	60
23	59	14	15	23	21	-10	-06	50
24	53	-27	-16	-09	00	13	18	44
25	29	13	16	09	31	-14	11	26

*The decimal point has been omitted for all entries.

TABLE 4
The Final Transformation Matrix*

	A	B	C	D	E	F	G
I	62	-01	-01	02	01	02	31
II	-53	82	-07	00	01	01	17
III	-20	-17	91	-11	01	27	-11
IV	-23	-12	-13	88	02	34	10
V	03	-06	-23	-28	85	36	-53
VI	28	45	21	30	24	-62	-35
VII	38	28	21	-20	-47	54	-67

*The decimal point has been omitted for all entries.

TABLE 5
The Oblique Factor Matrix*

Tests	A	B	C	D	E	F	G
1	—05	—05	36	—02	—02	46	07
2	00	02	44	—08	—01	—02	09
3	19	04	16	15	07	02	25
4	34	11	00	19	20	21	03
5	—05	02	—01	00	87	—05	01
6	14	02	—02	68	05	33	01
7	—04	—03	08	59	—01	—08	34
8	01	10	—05	02	—06	06	41
9	17	—02	19	17	05	45	09
10	30	—09	02	01	16	33	02
11	11	00	05	06	01	00	41
12	32	10	01	15	10	06	11
13	20	00	55	—04	10	—06	03
14	00	52	08	08	—06	—06	11
15	16	04	—05	—04	05	03	38
16	28	14	—05	18	09	16	07
17	18	09	18	27	03	09	26
18	14	01	09	09	36	05	09
19	—07	51	—11	—05	11	01	—01
20	25	05	10	15	04	03	16
21	20	26	38	08	—06	10	—08
22	13	14	—06	32	—10	39	12
23	16	—02	01	12	20	24	18
24	63	—08	—06	—05	—05	—05	—03
25	07	01	05	—09	19	34	—09

*The decimal point has been omitted for all entries.

TABLE 6
Correlations between the Primary Vectors

	A	B	C	D	E	F	G
A	1.00	.33	.16	.12	.29	.33	.43
B	.33	1.00	.16	.01	.27	.38	.40
C	.16	.16	1.00	.10	.35	.11	.32
D	.12	.01	.10	1.00	.07	.08	—14
E	.29	.27	.35	.07	1.00	.28	.41
F	.33	.38	.11	.08	.28	1.00	.47
G	.43	.40	.32	—14	.41	.47	1.00

TABLE 7
Loadings of the Primaries in the
Centroids of the Second Order

	I'	II'	III'
A	.57	.09	-.13
B	.53	.19	-.05
C	.43	-.36	.23
D	.12	-.27	-.35
E	.59	-.21	.14
F	.58	.24	-.13
G	.73	.28	.31

TABLE 8

	α	β	γ
I'	.38	.40	.66
II'	-.60	-.54	.71
III'	-.70	.74	-.26

TABLE 9
Rotated Factorial Matrix for the
Primaries in the Second Order

	α	β	γ
A	.25	.08	.47
B	.12	.07	.50
C	.22	.54	-.03
D	.45	-.07	-.02
E	.25	.45	.20
F	.17	.01	.59
G	-.11	.37	.60

TABLE 10
Correlations between the
Second-Order Factors

	α	β	γ
α	1.00	.04	.00
β	.04	1.00	.30
γ	.00	.30	.00

TABLE 11
Loadings of the Tests in the Orthogonal Second Order Factors

Test	I'	II'	III'	Test	I'	II'	III'
1	.48	.01	.06	14	.39	.10	.00
2	.28	-.12	.18	15	.53	.21	.14
3	.55	.02	.06	16	.51	.09	-.11
4	.62	.05	-.12	17	.62	.02	.00
5	.55	-.19	.16	18	.51	-.07	.05
6	.44	-.07	-.32	19	.28	.15	-.02
7	.33	-.10	-.04	20	.46	.02	.00
8	.45	.24	.13	21	.42	-.08	-.04
9	.64	.05	-.07	22	.50	.17	-.19
10	.49	.08	-.06	23	.59	.08	.00
11	.52	.14	.15	24	.22	.08	-.10
12	.51	.06	-.06	25	.34	.04	-.03
13	.45	-.21	.16				

REFERENCES

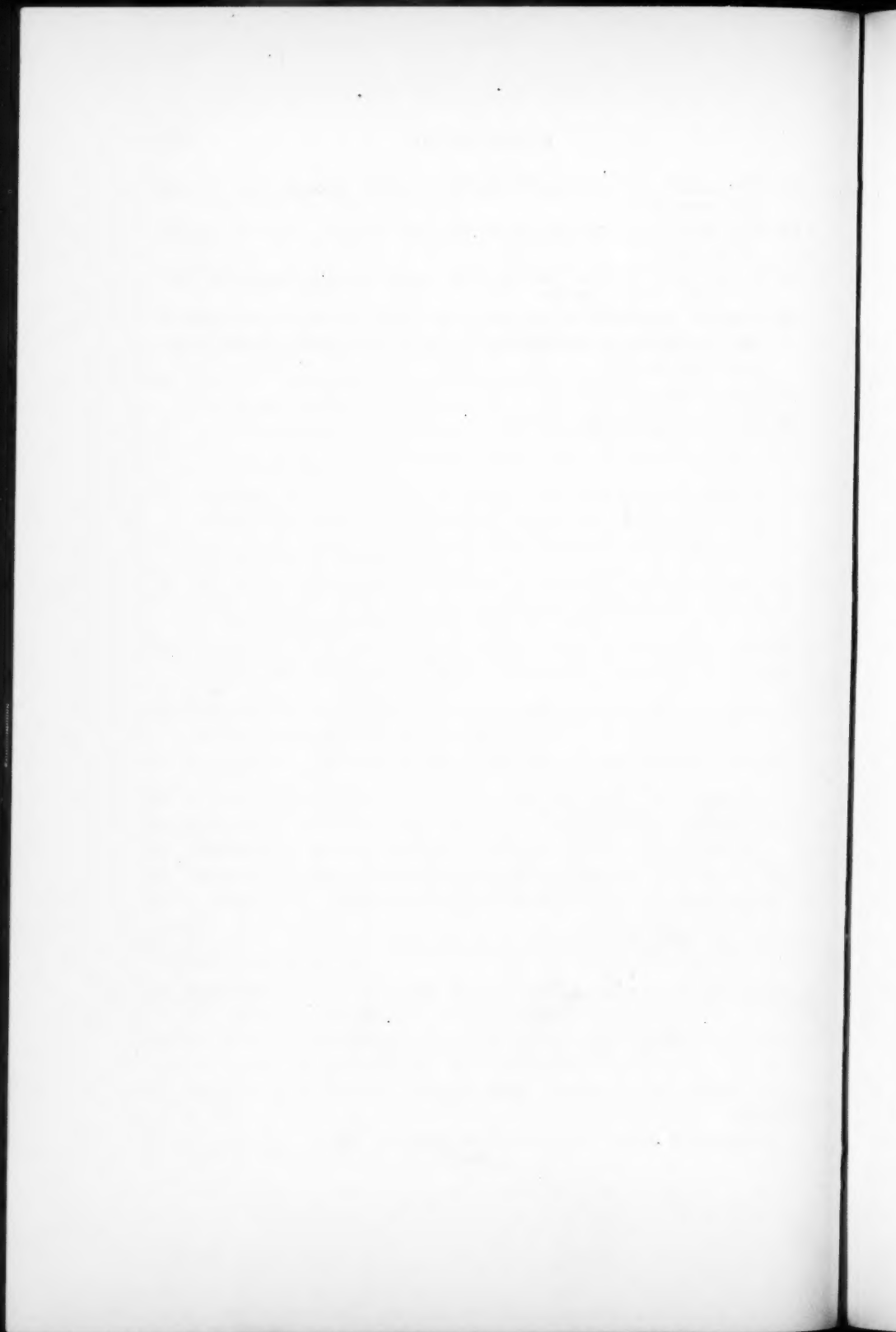
1. Alexander, W. P. Intelligence, concrete and abstract. *Brit. J. Psychol.*, Monograph Supplements, 1935, 6, No. 19.
2. Balinsky, B. An analysis of the mental factors of various age groups from nine to sixty. *Genet. Psychol. Monogr.*, 1942, 23, 191-234.
3. Bechtoldt, H. P. Factorial study of perceptual speed. Unpublished Ph.D. dissertation. University of Chicago, 1947.
4. Blakey, R. A re-analysis of a test of the theory of two factors. *Psychometrika*, 1940, 5, 121-135.
5. Blakey, R. A factor analysis of a non-verbal reasoning test. *Educ. psychol. Meas.*, 1941, 1, 187-198.
6. Brown, W. The mathematical and experimental evidence for the existence of a central intellectual factor. *Brit. J. Psychol.*, 1932-33, 23, 171-179.
7. Brown, W., and Stephenson, W. A test of the theory of two factors. *Brit. J. Psychol.*, 1932-33, 23, 352-370.
8. Burt, C. Mental and Scholastic Tests. London, 1922.
9. Burt, C. The Factors of the Mind. New York: The Macmillan Company, 1941.
10. Burt, C., and John, E. A factorial analysis of the Terman Binet tests. Part I and Part II. *Brit. J. educ. Psychol.*, 1942, 12, 156-161, 117-127.
11. Burt, C. Mental abilities and mental factors. *Brit. J. educ. Psychol.*, 1944, 14, 85-94.
12. Carroll, J. B. A factor analysis of verbal abilities. *Psychometrika*, 1941, 6, 279-308.
13. El Koussy, A. A. H. An investigation into the factors in tests involving the visual perception of space. *Brit. J. Psychol.*, Monograph Supplements, 1935, 7, No. 20, 1-87.

14. Goodman, C. H. Factorial analysis of Thurstone's seven primary abilities. *Psychometrika*, 1943, 8, 121-129.
15. Guilford, J. P. A note on the discovery of a *G* factor by means of Thurstone's centroid method of analysis. *Psychometrika*, 1941, 6, 205-207.
16. Hart, B., and Spearman, C. General ability, its existence and nature. *Brit. J. Psychol.*, 1912-13, 5, 51-84.
17. Holzinger, K. J. Preliminary report on Spearman-Holzinger Unitary Trait Study. No. 1 to No. 9. Prepared at the Statistical Laboratory, Dept. of Education, The University of Chicago.
18. Line, W. The growth of visual perception in children. *Brit. J. Psychol.*, Monograph Supplements, 1931, No. 15.
19. Mieli, R. L'analyse de l'intelligence. *Arch. Psychol.*, Geneve, 1943, 31, 1-64.
20. Mellone, M. A. A factorial study of picture tests for young children. *Brit. J. Psychol.*, 1944, 35, 9-16.
21. Moynihan, J. F. The concept of the synthetic sense and a technique of its measurement. *Stud. Psychol. Psychiat. Cathol. Univ. Amer.*, 1942, 5, No. 5.
22. Reyburn, H. A. and Taylor, J. C. Some factors of intelligence. *Brit. J. Psychol.*, 1941, 31, 259-270.
23. Rimoldi, H., Buhner, L., San Martin, R., Cortada, N., and Velasco, E. Desarrollo intelectual entre los 11 y los 14 años. *Publicaciones del Instituto de Psicologia Experimental.*, 1945, 1, No. 4, 119-237.
24. Rimoldi, H. J. A. Study of some factors related to intelligence. *Psychometrika*, 1948, 13, 27-46, and: Ph.D. Dissertation, University of Chicago, 1948.
25. Spearman, C. The Nature of Intelligence and the Principles of Cognition. London: Macmillan and Co., Ltd., 1927.
26. Spearman, C. Disturbers of tetrad differences. *J. educ. Psychol.*, 1930, 21, 559-573.
27. Spearman, C. The Abilities of Man. London, Macmillan Co., 1932.
28. Spearman, C. Thurstone's work reworked. *J. educ. Psychol.*, 1939, 30, 1-16.
29. Spearman, C. How "*G*" can disappear. *Psychometrika*, 1941, 6, 353-354.
30. Spearman, C. Theory of general factor. *Brit. J. Psychol.*, 1946, 36, 117-131.
31. Stephenson, W. Tetrad differences for verbal subtests. *J. educ. Psychol.*, 1931, 22, 255-267.
32. Stephenson, W. Tetrad differences for non-verbal subtests. *J. educ. Psychol.*, 1931, 22, 167-185.
33. Stephenson, W. Tetrad differences for verbal subtests relative to non-verbal subtests. *J. educ. Psychol.*, 1931, 22, 334-350.
34. Swineford, F. Some comparisons of the multiple factor and bi-factor methods of analysis. *Psychometrika*, 1941, 6, 375-382.
35. Thurstone, L. L. Primary mental abilities. *Psychometric Monographs*, No. 1, 1938.
36. Thurstone, L. L., and Thurstone, T. G. Factorial studies of intelligence. *Psychometric Monographs*, No. 2, 1942.

37. Thurstone, L. L. A factorial Study of Perception. Chicago: Univ. of Chicago Press, 1944.
38. Thurstone, L. L. Multiple Factor Analysis. Chicago: Univ. of Chicago Press, 1947.
39. Wright, R. E. A factor analysis of the original Stanford-Binet scale. *Psychometrika*, 1939, 4, 209-220.
40. Yela, M. Application of the concept of simple structure to Alexander's data. *Publication of the Psychometric Laboratory*, The University of Chicago, 1948, No. 49.

Manuscript received 3/5/50

Revised manuscript received 6/13/50



ON THE STANDARD LENGTH OF A TEST*

MAX A. WOODBURY

INSTITUTE FOR ADVANCED STUDY
AND
UNIVERSITY OF MICHIGAN

(1) A new descriptive parameter for tests, the standard length, is defined and related to reliability, correlation, and validity by means of simplified versions of known formulas. (2) The standard error of measurement is found to be related in simple fashion to the amount of information in a test in the sense of R. A. Fisher. The amount of information is computable as the test length divided by the standard length of the test. (3) The invariant properties of the standard length of a test under changes in length are discussed and proved. Similar results for the correlation coefficient corrected for attenuation and the index of validity are indicated.

Introduction

In connection with another study the notion of the *standard length* of a test turned out to be a useful means of simplifying notation and clarifying proofs. This brief note is presented to introduce this new and possibly valuable notion. The standard length is related to the *information* of R. A. Fisher† through the variance of the errors of measurement. There is an indirect relation to the type of information considered by Shannon‡ and later by Wiener.

It has long been recognized that the reliability of a test can be used (under certain restrictions which do not concern us here) to obtain the reliability of the test after it has been lengthened. Similar relations hold for the correlation between tests or the correlation of a test with a criterion (validity coefficient). This leads to the notion of the reliability and the validity as mathematical functions of the length of the test and the correlation between two tests as functions of the test lengths.

*The research covered by this note was supported by the Office of Naval Research.

†Fisher, R. A. *Statistical Methods for Research Workers*, 10th Edition. London: Oliver and Boyd, 1946, p. 346.

‡Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal*, 1948, 27, 379-423; 623-656.

Standard Length

The functional dependence of reliability upon test length is of a rather special algebraic character and involves only one parameter. It will be to the advantage of all if this parameter is chosen to simplify the formula. In the usual form of the relation the parameter is the reliability at a given (observed) length and gives the reliability for a test of e_i times the original length. The well known expression

$$\frac{e_i r_{ii}}{1 + (e_i - 1) r_{ii}} \quad (1)$$

expresses this relation.

If we rather arbitrarily define (however see the comment following (4)) the standard length of the test i as

$$\tau_i = \frac{t_i (1 - r_{ii})}{r_{ii}}, \quad (2)$$

where t_i is the observed test length and r_{ii} the observed reliability, then we find that the standard length computed for a test after it has been altered in length is the same as when computed for the original length. Specifically we find, since the new length is $e_i t_i$ and the new reliability is given in (1), that the new standard length is

$$\frac{(e_i t_i) \left[1 - \frac{e_i r_{ii}}{1 + (e_i - 1) r_{ii}} \right]}{\frac{e_i r_{ii}}{1 + (e_i - 1) r_{ii}}} = \frac{t_i (1 - r_{ii})}{r_{ii}}$$

which is the same as before. It should be noted in passing that any other invariant of the test must be a function of the standard length, where by an invariant of the test we mean any parameter of the test which does not depend on the length. It is clear that any invariant describes the contents of the test, not the accidental feature of its length. The standard length of a test together with its length determines the reliability. The formula for this purpose is

$$r_{ii} = \frac{t_i}{t_i + \tau_i}, \quad (3)$$

where t_i is the length of the test. From this it is easy to see that when a test has a length equal to its standard length it has a reli-

ability of one-half, when it has a length of twice its standard length it has a reliability of two-thirds, etc. As a matter of convenience we note that in order to obtain a reliability of r_{ii} the length of the test must be given by the relation

$$\frac{t_i}{\tau_i} = \frac{r_{ii}}{1 - r_{ii}}. \quad (4)$$

Other definitions of τ_i in (2) would lead to less simple formulas for (3) and (4) so that this may be considered as justification for the particular choice for τ_i .

Fisher has used a concept of information which gives the variance of errors as the reciprocal of the amount of information. This concept can be related to the reliability through the easily derived formula

$$r_{ii} = \frac{1}{1 + \sigma_i^2}, \quad (5)$$

where σ_i is the standard error of measurement and the standard deviation of the true scores is taken as a unit. Combining this equation with (3) we see that the amount of information is

$$J_i = \frac{1}{\sigma_i^2} = \frac{r_{ii}}{1 - r_{ii}} = \frac{t_i}{\tau_i}, \quad (6)$$

i.e., the length of the test measured in terms of its standard length as a unit. Thus a unit of information is the amount of information in a test of standard length.

Correlation and Validity

The formula analogous to (1) for computing the correlation between the tests i and j after each has been lengthened is*

$$r_{ij} \sqrt{\left(\frac{e_i}{1 + (e_i - 1)r_{ii}} \right) \left(\frac{e_j}{1 + (e_j - 1)r_{jj}} \right)} \quad (7)$$

where e_i and e_j are the ratios of the lengths of the lengthened tests to the original tests, and r_{ii} , r_{jj} and r_{ij} are the original reliabilities and correlation. By noting the relationship of (7) to (1) and (3) one can write down immediately the equation for the correlation as

*See Peters, C. C., and Van Voorhis, W. R. *Statistical Procedures and their Mathematical Bases*. N. Y.: McGraw-Hill Book Co., 1940, Eq. 111, p. 193.

a function of the lengths of the tests, viz.

$$r_{ij} = r_{i\infty, j\infty} \sqrt{\left(\frac{t_i}{t_i + \tau_i}\right) \left(\frac{t_j}{t_j + \tau_j}\right)} \quad (8)$$

where

$$r_{i\infty, j\infty} = \frac{r_{ij}}{\sqrt{r_{ii} r_{jj}}} \quad (9)$$

is the correlation coefficient corrected for attenuation and where the other symbols are defined as in (3). It should be noted that the correlation coefficient corrected for attenuation is the same for the lengthened tests as for the original tests so that it, like the standard lengths, is invariant under changes of length and describes a property of the content of the tests only and not of their lengths. To prove this, substitute in (9) from (7) and (1) to obtain the new coefficient corrected for attenuation. Further, any other invariant of the two tests must be a function of the three already described, viz., the standard lengths and the correlation coefficient corrected for attenuation.

The case of correlation with a criterion (validity) scarcely needs separate treatment. Let c denote the criterion and r_{ic} the validity coefficient of test i at length t_i and we have

$$r_{ic} = r_{i\infty, c} \sqrt{\frac{t_i}{t_i + \tau_i}} \quad (10)$$

where $r_{i\infty, c}$ is the index of validity, computable from the formula

$$r_{i\infty, c} = \frac{r_{ic}}{\sqrt{r_{ii}}} \quad (11)$$

The index of validity, like the standard length and the correlation corrected for attenuation, is invariant under changes in the length of the test i . From (10) we can find the length of the test which will give a specified validity: Note that only validities smaller in absolute value than the index of validity can be obtained and that the sign of the validity is unchanged by lengthening the test. Let r_{ic} be the desired validity, let t_i be the length of the test which will give this validity and we have

$$t_i = \frac{\tau_i r_{ic}^2}{r_{i\infty, c}^2 - r_{ic}^2} \quad (12)$$

Manuscript received 6/5/50

Revised Manuscript received 8/8/50

THE ESTIMATION OF THE PARAMETERS OF A NEGATIVE BINOMIAL DISTRIBUTION WITH SPECIAL REF- ERENCE TO PSYCHOLOGICAL DATA*

HERBERT S. SICHEL

NATIONAL INSTITUTE FOR PERSONNEL RESEARCH,
SOUTH AFRICAN COUNCIL FOR SCIENTIFIC AND INDUSTRIAL RESEARCH

As an analytical tool the negative binomial distribution may have wide applications in the psychological field. The estimation of its parameters is demonstrated to be often inefficient when fitting by the method of moments. This causes possibly true hypotheses to be rejected. Formulas for the efficiency of the moment method and solution of the likelihood equations are derived. Efficiency

graphs and detailed tables for the $\lambda(r, \hat{p})$ function reduce the maximum-likelihood method to a minimum of computational labour. Practical applications of the ease and power of the M.L. procedure are given.

Introduction

To mathematical statisticians the principles of maximum likelihood have always been a topic ever since R. A. Fisher (1) wrote his classic paper in 1921. On the applied side, however, the method has only come into vogue within the past decade. Of late, an ever-increasing stream of scientific papers appearing in a wide variety of journals testifies to the importance of "efficient estimation" in the practical field.

In the majority of psychometric studies we deal with normal, nearly normal, or normalized variables. This explains why the need for "efficient estimation" has not arisen in this branch of applied science to the extent that it has in others. It is a well-known fact, of course, that the customary moment statistics are efficient in the case of a normal population.

*The author wishes to express his gratitude to the South African Council for Scientific and Industrial Research for permission to publish this paper; to Mr. A. G. Arbous for supplying the data on absenteeism; to Mr. R. V. Sutton for the data on Two-hand Co-ordination errors; and last, but not least, to the staff of the Statistical Section who computed various tables under the expert guidance of Mr. J. S. Maritz.

On the other hand there exists one class of psychometric problems which can be measured only on a discrete scale. If the distribution of the discrete variable is very skew (J-shaped) it is virtually impossible to normalize the original data. At the best the transformation of a naturally discrete variable into a continuous one remains a dubious procedure. It appears to be more justified both from the theoretical and practical point of view, to work with the observed discrete variables.

A discrete probability distribution which is very flexible is the negative binomial law, sometimes referred to as the Greenwood-Yule curve (2). Its rationale was primarily built around the unequal liability of individuals to accidents and it may be derived as follows:

Suppose an individual has r accidents in a given time unit. If his exposure period is composed of many equal time units he will incur r accidents in $\gamma(r)$ units where $r = 0, 1, 2, 3 \dots k$. Provided the frequency distribution of his accidents may satisfactorily be described by the Poisson law we then may define the probability of his having r accidents per time unit as

$$\text{prob.}(r) = \frac{e^{-\lambda} \lambda^r}{r!}.$$

λ , the average number of accidents per time unit, may be considered as a measure of the individual's liability to accidents. In a group of operators, of whom all are exposed to the same environmental hazard, the parameter λ will vary from person to person according to the differential proneness to accidents. If the probability distribution of liabilities among the group is given as

$$dF = \frac{c^p}{\Gamma(p)} e^{-c\lambda} \lambda^{p-1} d\lambda, \quad (0 \leq \lambda < \infty)$$

we shall arrive at the number of persons having r accidents in a single time unit by integration of

$$\begin{aligned} \Lambda(r) &= N \int_0^\infty \frac{c^p}{\Gamma(p)} e^{-c\lambda} \lambda^{p-1} e^{-\lambda} \frac{\lambda^r}{r!} d\lambda \\ &= N \left(\frac{c}{c+1} \right)^p \frac{1}{\Gamma(p)} \frac{\Gamma(r+p)}{\Gamma(r+1)} \frac{1}{(c+1)^r}. \end{aligned}$$

The frequencies of 0, 1, 2, 3, accidents in the group are given, therefore, by

$$N \left(\frac{c}{c+1} \right)^p \left[1, \frac{p}{c+1}, \frac{p(p+1)}{2!(c+1)^2}, \frac{p(p+1)(p+2)}{3!(c+1)^3}, \dots \right]$$

which is the customary expression for the Greenwood-Yule distribution.

The negative binomial law may have wide and important applications in the psychological field, as it often is capable of describing satisfactorily such diverse phenomena as accidents, number of absences, number of days lost, errors made in a test, etc.

The two parameters of the negative binomial law are estimated, almost universally, by the method of moments. Even such an advanced mathematical statistics text as Kendall's (3) only refers to this procedure. Research workers on industrial accidents, without exception, use the same method. The only papers which draw attention to the inefficiency of the moment estimation in certain cases of the negative binomial law are Fisher's (4) and Haldane's (5) in *Annals of Eugenics*, a journal not ordinarily read by psychologists. The first part of the present investigation was undertaken by the author without any prior knowledge of Fisher's and Haldane's work. It, therefore, slightly differs from their approach but leads to identical results.

It will be shown that the practical application of the negative binomial law to phenomena in which the psychologist is mainly interested results very often in a rejection of a hypothesis which may really be true. Frequently, this is solely due to the employment of an inefficient estimation procedure.

The likelihood equation leads to mathematical expressions awkward to handle in the computation process. For this reason efficiency graphs and tables involving di- and trigamma functions were calculated and are included in this paper to facilitate the estimation of the unknown parameters. All the research worker now needs is an ordinary table of logarithms. The writer claims that, with the help of the graphs and tables, the computational labour has been cut down to a minimum. The advantage of using an efficient method of estimation is great, as illustrated by the examples quoted at the end of this paper.

The first four population moments and the large sample moment estimators of the negative binomial distribution

For reasons which will become apparent at a later stage, it is convenient to write the negative binomial law as

$$f(r) = \left(\frac{p}{p+m} \right)^p \frac{1}{\Gamma(p)} \frac{\Gamma(r+p)}{\Gamma(r+1)} \left(\frac{m}{p+m} \right)^r, \quad (1)$$

where $r = 0, 1, 2, \dots, \infty$, and m and p are the parameters in the law.

The characteristic function of (1) is

$$\phi(t) = \sum_{r=0}^{\infty} e^{itr} f(r) = \left[1 - (e^{it} - 1) \frac{m}{p} \right]^{-p},$$

and its cumulant generating function

$$\psi(t) = \log \phi(t) = -p \log \left[1 - (e^{it} - 1) \frac{m}{p} \right]. \quad (2)$$

Expansion of (2) gives

$$\begin{aligned} \psi(t) = p \left\{ \left[\frac{it}{1!} + \frac{(it)^2}{2!} + \frac{(it)^3}{3!} + \frac{(it)^4}{4!} \dots \right] \left(\frac{m}{p} \right) \right. \\ \left. + \frac{1}{2} \left[\frac{it}{1!} + \frac{(it)^2}{2!} + \frac{(it)^3}{3!} \dots \right]^2 \left(\frac{m}{p} \right)^2 \right. \\ \left. + \frac{1}{6} \left[\frac{it}{1!} + \frac{(it)^2}{2!} \dots \right]^3 \left(\frac{m}{p} \right)^3 + \frac{1}{24} \left[\frac{it}{1!} \dots \right]^4 \left(\frac{m}{p} \right)^4 + \dots \right\}. \quad (3) \end{aligned}$$

By expanding the square brackets in (3) and collecting terms of $\frac{(it)^j}{j!}$ we find the j th cumulant being the factor of $\frac{(it)^j}{j!}$. We then have

$$\left. \begin{aligned} \kappa_1 &= m, \\ \kappa_2 &= m + \frac{m^2}{p}, \\ \kappa_3 &= m + \frac{3m^2}{p} + \frac{2m^3}{p^2}, \text{ and} \\ \kappa_4 &= m + \frac{7m^2}{p} + \frac{12m^3}{p^2} + \frac{6m^4}{p^3}. \end{aligned} \right\} \quad (4)$$

Finally, from the well known relations between cumulants and moments,

$$\mu_1' = m, \quad (5a)$$

$$\mu_2 = m + \frac{m^2}{p}, \quad (5b)$$

$$\mu_3 = m + \frac{3m^2}{p} + \frac{2m^3}{p^2}, \text{ and} \quad (5c)$$

$$\mu_4 = m + 3m^2 + \frac{7m^2 + 6m^3}{p} + \frac{12m^3 + 3m^4}{p^2} + \frac{6m^4}{p^3}. \quad (5d)$$

One of the advantages of choosing the law in the form of equation (1) is brought out readily by equation (5a): The parameter m is identical with the population mean μ_1' .

From (5a) and (5b) we find the moment estimators \bar{m} and \bar{p} of the parameters m and p , for large sample size n ,

$$\bar{m} = m_1', \text{ and} \quad (6a)$$

$$\bar{p} = \frac{m_1'^2}{m_2 - m_1'}, \quad (6b)$$

where m_1' and m_2 the estimates of the parent moments μ_1' and μ_2 .

The variances of moment estimators \bar{m} and \bar{p} for large samples
The variance of the mean is

$$\text{Var} (m_1') = \frac{\mu_2}{n},$$

and, by substitution of (5b) into the above,

$$\text{Var} (\bar{m}) = \text{Var} (m_1') = \frac{m(p + m)}{pn}. \quad (7)$$

\bar{p} is a function of the moments, say

$$\bar{p} = \phi(m_1', m_2),$$

hence, for large n , approximately

$$\begin{aligned} \text{Var} (\bar{p}) = & \left(\frac{\partial \phi}{\partial m_1'} \right)^2 \text{Var} (m_1') + 2 \frac{\partial \phi}{\partial m_1'} \frac{\partial \phi}{\partial m_2} \text{Cov} (m_1', m_2) \\ & + \left(\frac{\partial \phi}{\partial m_2} \right)^2 \text{Var} (m_2). \end{aligned} \quad (8)$$

Substitution of the appropriate partial derivatives of (6b) and of

standard expressions for $\text{Var}(m_1')$, $\text{Var}(m_2)$ and $\text{Cov}(m_1', m_2)$ into equation (8) gives (after replacing the statistics by their expectations)

$$\text{Var}(\bar{p}) = \frac{\mu_1'^2}{n(\mu_2 - \mu_1')^4} \times [(2\mu_2 - \mu_1')^2 \mu_2 + \mu_1'^2(\mu_4 - \mu_2^2) - 2\mu_1' \mu_3(2\mu_2 - \mu_1')]. \quad (9)$$

Substitution of (5a, b, c, d) into (9) yields

$$\text{Var}(\bar{p}) = \frac{2p(1+p)(1+p/m)^2}{n}, \quad (10)$$

this being the large sample variance of \bar{p} .

The efficiency of moment estimators \bar{m} and \bar{p} for large samples

If we want to gauge the efficiency of the moment estimators we must evaluate the Hessian determinant

$$\Delta = \begin{vmatrix} -E\left(\frac{\partial^2 \log f(r)}{\partial m^2}\right) & -E\left(\frac{\partial^2 \log f(r)}{\partial m \partial p}\right) \\ -E\left(\frac{\partial^2 \log f(r)}{\partial m \partial p}\right) & -E\left(\frac{\partial^2 \log f(r)}{\partial p^2}\right) \end{vmatrix}, \quad (11)$$

whence the variances of the maximum-likelihood estimators \hat{m} and \hat{p} (for large n) are

$$\text{Var}(\hat{m}) = -\frac{1}{\Delta n} E\left(\frac{\partial^2 \log f(r)}{\partial p^2}\right), \quad \text{and} \quad (12)$$

$$\text{Var}(\hat{p}) = -\frac{1}{\Delta n} E\left(\frac{\partial^2 \log f(r)}{\partial m^2}\right). \quad (13)$$

Now

$$\frac{\partial^2 \log f(r)}{\partial m \partial p} = \frac{r-m}{(p+m)^2},$$

and its expected mean value

$$E\left(\frac{\partial^2 \log f(r)}{\partial m \partial p}\right) = \sum_{r=0}^{\infty} \frac{r-m}{(p+m)^2} f(r) = 0.$$

Hence it follows that \hat{m} and \hat{p} are uncorrelated. This is the second advantage of writing the negative binomial in the form of equation (1). Equations (12) and (13) simplify into

$$\text{Var}(\hat{m}) = -\frac{1}{E\left(\frac{\partial^2 \log f(r)}{\partial m^2}\right)n}, \quad (12a)$$

$$\text{Var}(\hat{p}) = -\frac{1}{E\left(\frac{\partial^2 \log f(r)}{\partial p^2}\right)n}. \quad (13a)$$

Writing

$$\text{Trig.}(x) = \frac{d^2 \log \Gamma(x+1)}{dx^2},$$

where $\text{Trig.}(x)$ is the trigamma function, we find after some differentiation and summation

$$\text{Var}(\hat{m}) = \frac{m(p+m)}{np}, \quad (14)$$

$$\begin{aligned} \text{Var}(\hat{p}) = 1 / & n \left[\left(\frac{p}{p+m} \right)^p \frac{1}{\Gamma(p)} \sum_{r=0}^{\infty} \left(\frac{m}{p+m} \right)^r \frac{\Gamma(r+p)}{\Gamma(r+1)} \right. \\ & \times \left. \left\{ \text{Trig.}(x)(p-1) - \text{Trig.}(x)(p+r-1) \right\} - \frac{m}{p(p+m)} \right]. \end{aligned} \quad (15)$$

Equations (7) and (14) are identical. Hence it follows that the moment estimator \hat{m} has maximum efficiency. It may also be shown that \hat{m} is a sufficient estimator. The efficiency of \hat{p} is

$$\text{Eff.}(\hat{p}) = \frac{\text{Var.}(\hat{p})}{\text{Var.}(\bar{p})} = \frac{\text{eq. (15)}}{\text{eq. (10)}}. \quad (16)$$

Equations (15) and (16) may be simplified considerably, as indicated by Fisher (4).

By writing

$$\frac{p}{p+m} = 1-c,$$

$$\frac{m}{p+m} = c,$$

and using the relation

$$\text{Trig.}(x)(p-1) - \text{Trig.}(x)(p+r-1) = \sum_{v=1}^r (p+v-1)^2,$$

where $\text{Trig.}(x)$ is the trigamma function, we find, after some algebra,

$$\text{Var.}(\hat{p}) = \left[n \sum_{r=2}^{\infty} \frac{1}{r} \left(\frac{m}{p+m} \right)^r B(r, p) \right]^{-1} \quad (15a)$$

and

$$\text{Eff.}(\bar{p}) = \left[2 \sum_{r=2}^{\infty} \frac{1}{r} \left(\frac{m}{p+m} \right)^{r-2} \frac{\Gamma(r)\Gamma(p+2)}{\Gamma(r+p)} \right]^{-1} \quad (16a)$$

The reciprocal of the efficiency, expanded as a series, gives

$$[\text{Eff.}(\bar{p})]^{-1} = 1 + \frac{2 \times 2!}{3(p+2)} \left(\frac{m}{p+m} \right) + \frac{2 \times 3!}{4(p+2)(p+3)} \left(\frac{m}{p+m} \right)^2$$

$$+ \frac{2 \times 4!}{5(p+2)(p+3)(p+4)} \left(\frac{m}{p+m} \right)^3 + \dots \quad (17)$$

which is the same as Fisher's expression (4).

For p small and finite m (17) tends towards the divergent series

$$2\left(\frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots\right);$$

hence

$$\lim_{p \rightarrow 0} \text{Eff.}(\bar{p}) = 0.$$

For $p \rightarrow \infty$ and all values of m ,

$$\lim_{p \rightarrow \infty} \text{Eff.}(\bar{p}) = 1.$$

For $m \rightarrow 0$ and p finite,

$$\lim_{m \rightarrow 0} \text{Eff.}(\bar{p}) = 1.$$

For m large, equations (15) and (10) become

$$\text{Var.}(\hat{p}) \approx \frac{p}{n[p \text{ Trig.}(x)(p-1) - 1]}$$

$$\text{Var.}(\bar{p}) \approx \frac{2p(1+p)}{n};$$

hence

$$\lim_{m \rightarrow \infty} \text{Eff.}(\bar{p}) = \{2(p+1)[p \text{ Trig.}(x)(p-1) - 1]\}^{-1}, \quad (18)$$

where $\text{Trig.}(x)$ is the trigamma function.

Equation (18) gives the lower bound of $\text{Eff.}(\bar{p})$ below which the efficiency of estimator \bar{p} cannot sink. For $p = 5.5$ the minimum efficiency of moment estimator \bar{p} is .80. For values of $p \geq 5.5$ it is not necessary, therefore, to estimate p by the more arduous maximum-likelihood method.

Exact values of $\text{Eff.}(\bar{p})$ were calculated for various levels of p and m . They are graphed in Figure 1. After a preliminary estimation of the parameters by the method of moments, Figure 1 will indicate whether it is advisable to estimate by maximum likelihood. Although we enter Figure 1 with the (often) inefficient moment estimate, it has been found in practice that the decision arrived at from the graph is almost always right. This may be verified by subsequently reading off the efficiency based on the maximum-likelihood estimates and comparing with the efficiency value found from the moment solution. It is desirable to estimate by the method of maximum likelihood whenever $\text{Eff.}(\bar{p}) < .80$.

Large sample maximum-likelihood estimators \hat{m} and \hat{p}

The likelihood solution is given by

$$\begin{aligned} \frac{\partial \log L}{\partial m} &= -\frac{p}{p+m}n + \frac{p}{m(p+m)}\sum r = 0, \\ \frac{\partial \log L}{\partial p} &= n \log \left(\frac{p}{p+m} \right) - n \text{Dig.}(x)(p-1) \\ &\quad + \frac{m}{p+m}n - \frac{1}{p+m}\sum r + \sum \text{Dig.}(x)(p+r-1) = 0, \end{aligned} \quad (19)$$

where we write

$$\text{Dig.}(x) = \frac{d \log \Gamma(x+1)}{dx},$$

where $\text{Dig.}(x)$ is the digamma function. By solving (19) for \hat{m} and \hat{p} we find

$$\hat{m} = \frac{1}{n} \sum_{i=1}^n r_i, \quad (20)$$

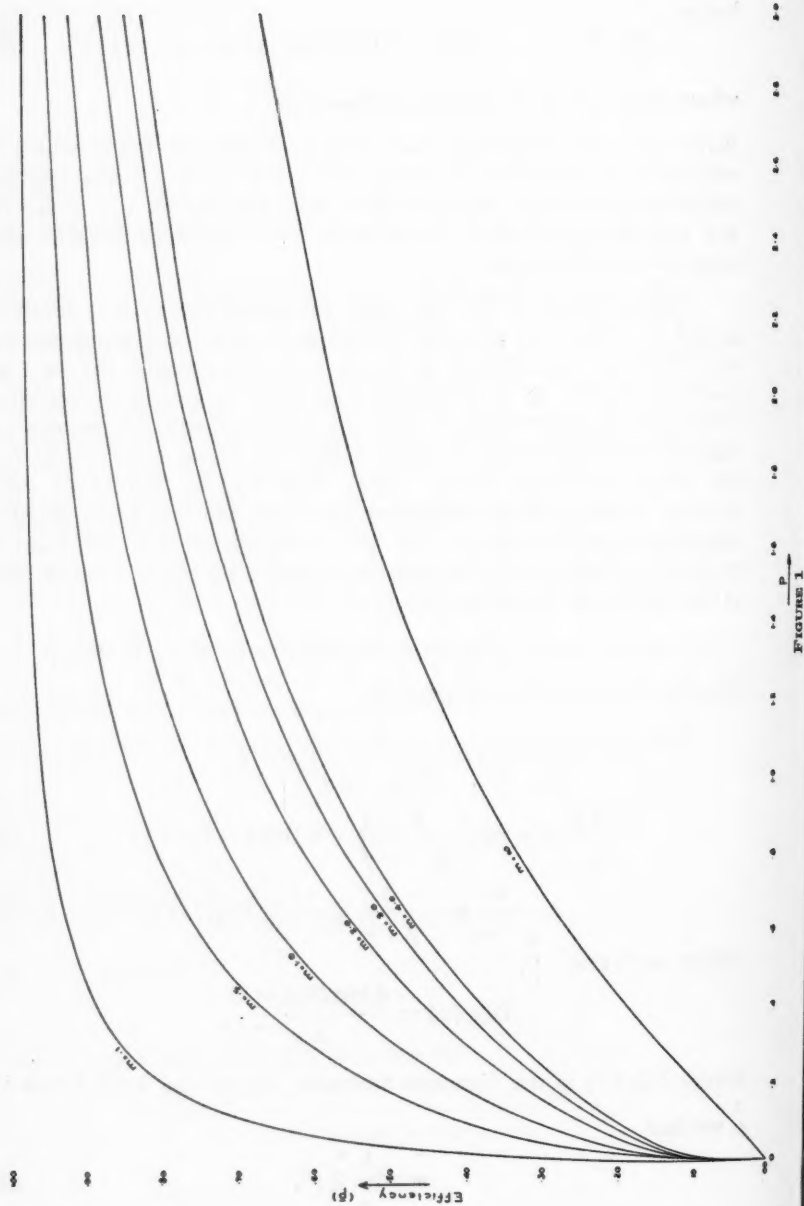
EFFICIENCY OF MOMENT ESTIMATOR \hat{p} OF NEGATIVE BINOMIAL DISTRIBUTION

FIGURE 1

$$-\sum_{n=1}^{\infty} [\text{Dig.}(x) (\hat{p} + r, -1) - \text{Dig.}(x) (\hat{p} - 1)] - \log \left(1 + \frac{\hat{m}}{\hat{p}} \right) = 0, \quad (21)$$

where $\text{Dig.}(x)$ is the digamma function.

To facilitate the numerical solution of equation (21) for \hat{p} the function

$$\lambda(r, \hat{p}) = \text{Dig.}(x) (\hat{p} + r - 1) - \text{Dig.}(x) (\hat{p} - 1)$$

has been tabulated for various values of \hat{p} and for $r = 0, 1, 2, \dots, 35$.

The table of $\lambda(r, \hat{p})$ is included in this paper (Table 1). Should values of $r > 35$ be required we may make use of a simple approximation for large r , i.e.

$$\begin{aligned} \lambda(r, \hat{p}) &\approx \log(\hat{p} + r - 1) \\ &+ \frac{1}{2(\hat{p} + r - 1)} - \frac{1}{12(\hat{p} + r - 1)^2} - \text{Dig.}(x) (\hat{p} - 1). \end{aligned}$$

The function $\text{Dig.}(x) (\hat{p} - 1)$ is given in Table 2.

The variances of the maximum-likelihood estimators \hat{m} and \hat{p} were previously given as equations (14) and (15a).

Applications

(a) Absence Proneness:

In an intensive study of absenteeism among industrial workers, the result of which is to be published in due course, it was established that liability to absence from work differs from person to person. Thus we may speak of absence proneness just as we also speak of accident proneness. In order to test for absence proneness we may collect data on the number of absences of individuals in a given time unit, arrange them into a frequency distribution, and fit the negative binomial. If the resulting χ^2 -test is satisfactory we may take this as an indication that among the human population from which the sample was drawn there exists unequal liability to absenteeism. (This procedure of testing the hypothesis is a necessary one but it is not sufficient, as shown by Maritz (6).)

Now Fisher (7) has shown that the use of the χ^2 -test is only legitimate if the method of estimation is "efficient." In the case of

TABLE 1
The Function $\lambda(\tau, p)^*$

$\tau \backslash \hat{p}$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	0	0	0	0	0	0	0	0	0	0
1	10.00000	5.00000	3.33333	2.50000	2.00000	1.66667	1.42857	1.25000	1.11111	1.00000
2	10.90909	5.83333	4.10256	3.21428	2.66667	2.29167	2.01680	1.80556	1.63743	1.50000
3	11.38528	6.28788	4.53734	3.63095	3.06667	2.67628	2.38717	2.16270	1.98226	1.83334
4	11.70786	6.60038	4.84037	3.92506	3.35238	2.95406	2.65744	2.42586	2.23867	2.08334
5	11.95176	6.83847	5.07293	4.15234	3.57460	3.17145	2.87021	2.63419	2.44275	2.28334
6	12.14784	7.03078	5.26161	4.33752	3.75642	3.35002	3.04565	2.80660	2.61224	2.45000
7	12.31177	7.19207	5.42034	4.49377	3.91027	3.50154	3.19490	2.95366	2.75717	2.59286
8	12.45262	7.33096	5.55733	4.62891	4.04360	3.63312	3.32477	3.08187	2.88375	2.71786
9	12.57607	7.45291	5.67781	4.74796	4.16125	3.74940	3.43971	3.19551	2.99611	2.82897
10	12.68596	7.56161	5.78534	4.85434	4.26651	3.85356	3.54281	3.29755	3.09712	2.92897
11	12.78497	7.65965	5.88242	4.95049	4.36175	3.94790	3.63626	3.39014	3.18886	3.01988
12	12.87506	7.74893	5.97092	5.03821	4.44871	4.03411	3.72174	3.47488	3.27290	3.10322
13	12.95771	7.83090	6.05222	5.11886	4.52871	4.11347	3.80048	3.55301	3.35042	3.18014
14	13.03404	7.90666	6.12741	5.19348	4.60278	4.18700	3.87347	3.62547	3.42236	3.25157
15	13.10497	7.97708	6.19734	5.26293	4.67175	4.25550	3.94150	3.69304	3.48947	3.31823
16	13.17119	8.04287	6.26270	5.32786	4.73626	4.31960	4.00519	3.75633	3.55237	3.38073
17	13.23330	8.10460	6.32405	5.38884	4.79687	4.37984	4.06507	3.81585	3.61154	3.43956
18	13.29178	8.16274	6.38185	5.44631	4.85401	4.43666	4.12157	3.87203	3.66740	3.49511
19	13.34703	8.21768	6.43649	5.50066	4.90806	4.49042	4.17504	3.92523	3.72031	3.54774
20	13.39939	8.26976	6.48831	5.55220	4.95935	4.54144	4.22580	3.97573	3.77056	3.59774
21	13.44914	8.31927	6.53757	5.60122	5.00812	4.58998	4.27411	4.02381	3.81840	3.64536
22	13.49653	8.36644	6.58452	5.64795	5.05464	4.63628	4.32019	4.06968	3.86408	3.69082
23	13.54178	8.41148	6.62936	5.69259	5.09909	4.68053	4.36425	4.11354	3.90774	3.73429
24	13.58506	8.45459	6.67228	5.73533	5.14164	4.72290	4.40644	4.15555	3.94958	3.77596
25	13.62657	8.49591	6.71344	5.77631	5.18245	4.76355	4.44692	4.19587	3.98975	3.81597
26	13.66641	8.53559	6.75296	5.81568	5.22167	4.80262	4.48584	4.23464	4.02835	3.85442
27	13.70472	8.57376	6.79098	5.85356	5.25940	4.84021	4.52329	4.27195	4.06553	3.89146
28	13.74162	8.61053	6.82761	5.89006	5.29576	4.87644	4.55939	4.30792	4.10137	3.92718
29	13.77721	8.64599	6.86295	5.92527	5.33086	4.91141	4.59424	4.34265	4.13597	3.96166
30	13.81157	8.68022	6.89708	5.95928	5.36475	4.94519	4.62791	4.37620	4.16942	3.99499
31	13.84479	8.71335	6.93008	5.99218	5.39754	4.97787	4.66048	4.40867	4.20178	4.02725
32	13.87695	8.74540	6.96203	6.02403	5.42928	5.00951	4.69202	4.44012	4.23313	4.05850
33	13.90810	8.77645	6.99299	6.05489	5.46006	5.04019	4.72261	4.47061	4.26352	4.08880
34	13.93831	8.80657	7.02302	6.08483	5.48991	5.06995	4.75228	4.50019	4.29302	4.11821
35	13.96764	8.83582	7.05217	6.11390	5.51889	5.09886	4.78110	4.52893	4.32168	4.14679

*Due to rounding, tabular values may be in error by one in the last unit.

TABLE 1 (Continued)
The Function $\lambda(r, p)^*$

$\frac{r}{p}$	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	0	0	0	0	0	0	0	0	0	0
1	.90909	.83333	.76923	.71428	.66667	.62500	.58823	.55556	.52632	.50000
2	1.38528	1.28788	1.20401	1.13095	1.06667	1.00961	.95860	.91270	.87115	.83334
3	1.70786	1.60038	1.50704	1.42506	1.35238	1.28739	1.22887	1.17586	1.12756	1.08334
4	1.95176	1.83847	1.73960	1.65234	1.57460	1.50478	1.44164	1.38419	1.33164	1.28334
5	2.14784	2.03078	1.92828	1.83752	1.75642	1.68335	1.61708	1.55660	1.50113	1.45000
6	2.31177	2.19207	2.08701	1.99377	1.91027	1.83487	1.76633	1.70366	1.64606	1.59286
7	2.45262	2.33096	2.22400	2.12891	2.04360	1.96645	1.89620	1.83187	1.77264	1.71786
8	2.57607	2.45291	2.34448	2.24796	2.16125	2.08273	2.01114	1.94551	1.88500	1.82897
9	2.68596	2.56161	2.45201	2.35434	2.26651	2.18689	2.11424	2.04755	1.98601	1.92897
10	2.78497	2.65965	2.54909	2.45049	2.36175	2.28123	2.20769	2.14014	2.07775	2.01988
11	2.87506	2.74893	2.63759	2.53821	2.44871	2.36744	2.29317	2.22488	2.16179	2.10322
12	2.95771	2.83090	2.71889	2.61886	2.52871	2.44680	2.37191	2.30301	2.23931	2.18014
13	3.03404	2.90666	2.79408	2.69348	2.60278	2.52033	2.44490	2.37547	2.31125	2.25157
14	3.10497	2.97708	2.86401	2.76293	2.67175	2.58883	2.51293	2.44304	2.37836	2.31823
15	3.17119	3.04287	2.92937	2.82786	2.73626	2.65293	2.57662	2.50633	2.44126	2.38073
16	3.23330	3.10460	2.99072	2.88884	2.79687	2.71317	2.63650	2.56586	2.50043	2.43956
17	3.29178	3.16274	3.04852	2.94631	2.85401	2.76999	2.69300	2.62203	2.55629	2.49511
18	3.34703	3.21768	3.10316	3.00066	2.90806	2.82375	2.74647	2.67523	2.60920	2.54774
19	3.39939	3.26976	3.15498	3.05220	2.95935	2.87477	2.79723	2.72573	2.65945	2.59774
20	3.44914	3.31927	3.20424	3.10122	3.00812	2.92331	2.84554	2.77381	2.70729	2.64536
21	3.49653	3.36644	3.25119	3.14795	3.05464	2.96961	2.89162	2.81968	2.75297	2.69082
22	3.54178	3.41148	3.29603	3.19259	3.09909	3.01386	2.93568	2.86354	2.79663	2.73429
23	3.58506	3.45459	3.33895	3.23533	3.14164	3.05623	2.97787	2.90555	2.83847	2.77596
24	3.62657	3.49591	3.38011	3.27631	3.18245	3.09688	3.01835	2.94587	2.87864	2.81597
25	3.66641	3.53559	3.41963	3.31568	3.22167	3.13595	3.05727	2.98464	2.91724	2.85442
26	3.70472	3.57376	3.45765	3.35356	3.25940	3.17354	3.09472	3.02195	2.95442	2.89146
27	3.74162	3.61053	3.49428	3.39006	3.29576	3.20977	3.13082	3.05792	2.99026	2.92718
28	3.77721	3.64599	3.52962	3.42527	3.33086	3.24474	3.16567	3.09265	3.02486	2.96166
29	3.81157	3.68022	3.56375	3.45928	3.36475	3.27852	3.19934	3.12620	3.05831	2.99499
30	3.84479	3.71335	3.59675	3.49218	3.39754	3.31120	3.23191	3.15867	3.09067	3.02725
31	3.87695	3.74540	3.62870	3.52403	3.42928	3.34234	3.26345	3.19012	3.12202	3.05850
32	3.90810	3.77645	3.65966	3.55489	3.46006	3.37352	3.29404	3.22061	3.15241	3.08880
33	3.93831	3.80657	3.68969	3.58483	3.48991	3.40328	3.32371	3.25019	3.18191	3.11821
34	3.96764	3.83582	3.71884	3.61390	3.51889	3.43219	3.35253	3.27893	3.21057	3.14679
35	3.99613	3.86422	3.74717	3.64215	3.54706	3.46027	3.38054	3.30686	3.23842	3.17456

*Due to rounding, tabular values may be in error by one in the last unit.

TABLE 1 (Continued)

The Function $\lambda(r, \hat{p})^*$

$\hat{p} \backslash r$	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0
0	0	0	0	0	0	0	0	0	0	0
1	.47619	.45455	.43478	.41667	.40000	.38461	.37037	.35714	.34483	.33334
2	.79877	.76705	.73781	.71078	.68571	.66239	.64064	.62030	.60124	.58334
3	1.04267	1.00514	.97037	.93806	.90793	.87978	.85341	.82863	.80532	.78334
4	1.23875	1.19745	1.15905	1.12324	1.08975	1.05835	1.02885	1.00104	.97481	.95000
5	1.40268	1.35874	1.31778	1.27949	1.24360	1.20987	1.17810	1.14810	1.11974	1.09286
6	1.54353	1.49763	1.45477	1.41463	1.37693	1.34145	1.30797	1.27631	1.24632	1.21786
7	1.66698	1.61958	1.57525	1.53368	1.49458	1.45773	1.42291	1.38995	1.35868	1.32897
8	1.77687	1.72828	1.68278	1.64006	1.59984	1.56189	1.52601	1.49199	1.45969	1.42897
9	1.87588	1.82632	1.77986	1.73621	1.69508	1.65623	1.61946	1.58458	1.55143	1.51988
10	1.96597	1.91560	1.86836	1.82393	1.78204	1.74244	1.70494	1.66932	1.63547	1.60322
11	2.04862	1.99757	1.94966	1.90458	1.86204	1.82180	1.78368	1.74745	1.71299	1.68014
12	2.12495	2.07333	2.02485	1.97920	1.93611	1.89533	1.85667	1.81991	1.78493	1.75157
13	2.19588	2.14375	2.09478	2.04865	2.00508	1.96383	1.92470	1.88748	1.85204	1.81823
14	2.26210	2.20954	2.16014	2.11358	2.06959	2.02793	1.98839	1.95077	1.91494	1.88073
15	2.32421	2.27127	2.22149	2.17456	2.13020	2.08817	2.04827	2.01030	1.97411	1.93956
16	2.38269	2.32941	2.27929	2.23203	2.18734	2.14499	2.10477	2.06647	2.02997	1.99511
17	2.43794	2.38435	2.33393	2.28638	2.24139	2.19875	2.15824	2.11967	2.08288	2.04774
18	2.49030	2.43643	2.38575	2.33792	2.29268	2.24977	2.20900	2.17017	2.13313	2.09774
19	2.54005	2.48594	2.43501	2.38694	2.34145	2.29831	2.25731	2.21825	2.18097	2.14536
20	2.58744	2.53311	2.48196	2.43367	2.38797	2.34461	2.30339	2.26412	2.22665	2.19082
21	2.63269	2.57815	2.52680	2.47831	2.43242	2.38886	2.34745	2.30798	2.27031	2.23429
22	2.67597	2.62126	2.56972	2.52105	2.47497	2.43123	2.38964	2.34999	2.31215	2.27596
23	2.71748	2.66258	2.61088	2.56203	2.51578	2.47188	2.43012	2.39031	2.35232	2.31597
24	2.75732	2.70226	2.65040	2.60140	2.55500	2.51095	2.46904	2.42908	2.39092	2.35442
25	2.79563	2.74043	2.68842	2.63928	2.59273	2.54854	2.50649	2.46639	2.42810	2.39146
26	2.83253	2.77720	2.72505	2.67578	2.62909	2.58477	2.54259	2.50236	2.46394	2.42718
27	2.86812	2.81266	2.76039	2.71099	2.66419	2.61974	2.57744	2.53709	2.49854	2.46166
28	2.90248	2.84689	2.79452	2.74500	2.69808	2.65352	2.61111	2.57064	2.53199	2.49499
29	2.93570	2.88002	2.82752	2.77790	2.73087	2.68620	2.64368	2.60311	2.56435	2.52725
30	2.96786	2.91207	2.85947	2.80975	2.76261	2.71784	2.67522	2.63456	2.59570	2.55850
31	2.99901	2.94312	2.89043	2.84061	2.79339	2.74852	2.70581	2.66505	2.62609	2.58880
32	3.02922	2.97324	2.92046	2.87055	2.82324	2.77828	2.73548	2.69463	2.65559	2.61821
33	3.05855	3.00249	2.94961	2.89962	2.85222	2.80719	2.76430	2.72337	2.68425	2.64679
34	3.08704	3.03089	2.97794	2.92787	2.88039	2.83527	2.79231	2.75130	2.71210	2.67456
35	3.11474	3.05852	3.00549	2.95534	2.90779	2.86260	2.81956	2.77847	2.73920	2.70159

*Due to rounding, tabular values may be in error by one in the last unit.

TABLE 2
The Function $\text{Dig.}(x) (\hat{p} - 1)^*$

\hat{p}	$\text{Dig.}(x) (\hat{p} - 1)$	\hat{p}	$\text{Dig.}(x) (\hat{p} - 1)$
.1	— 10.42375	1.6	.12605
.2	— 5.28904	1.7	.20855
.3	— 3.50252	1.8	.28499
.4	— 2.56138	1.9	.35618
.5	— 1.96351	2.0	.42278
.6	— 1.54062	2.1	.48534
.7	— 1.22002	2.2	.54429
.8	— .96501	2.3	.60004
.9	— .75493	2.4	.65290
1.0	— .57722	2.5	.70316
1.1	— .42375	2.6	.75105
1.2	— .28904	2.7	.79678
1.3	— .16919	2.8	.84055
1.4	— .06138	2.9	.88250
1.5	+ .03649	3.0	.92278

*Due to rounding, tabular values may be in error by one in the last unit.

an inefficient estimation process it may happen that, on the strength of the χ^2 -test, we reject a hypothesis which may really be true.

The observed distribution of number of absences of 318 workers in a particular division of a large steel corporation is given in the first two columns of Table 3. The observational period was six months and the absences were all classified as "absent without permission." For the sample mean and variance we find

$$m_1' = .66981,$$

$$m_2 = 1.52934,$$

and hence from (6a) and (6b) the moment estimates

$$\bar{m} = .66981,$$

$$\bar{p} = .52197.$$

Substituting these estimates into the negative binomial law (eq. 1) and calculating expected frequencies, gives finally a total $\chi^2 = 7.188$ (columns 3-5 in Table 3). For such a value of χ^2 and two degrees of freedom, $P = .03$. We should therefore reject the hypothesis of absence proneness, as described by the Greenwood-Yule model (2).

However, as mentioned previously, the rejection of a true hypothesis may be entirely due to what Fisher calls "errors of estima-

TABLE 3

Comparison of the Method of Moments and the Method of Maximum Likelihood in Fitting the Negative Binomial in Studying Absence Proneness

Method of Estimation			Moments			Maximum Likelihood	
Number of Absences	Observed Frequency, f_o	Expected Frequency, f_E	$f_o - f_E$	χ^2	f'_E	$f_o - f'_E$	χ^2
0	217	206.7	+10.3	.513	214.9	+2.1	.0
1	44	60.6	-16.6	4.547	53.4	-9.4	1.6
2	29	25.9	+ 3.1	.371	23.4	+5.6	1.3
3	11	12.3	- 1.3	.137	11.8	-0.8	.0
4	11	6.1	+ 4.5	1.620	6.3	+2.5	.4
5	2	3.1			3.5		
6	4	1.6			2.0		
7 & Over	0	1.7			2.7		
Totals	318	318.0	0.0	7.188	318.0	0.0	3.5
			D.F. = 2		P = .03 D.F. = 2		P =

tion" if we fit with an inefficient method. By entering Figure 1 with the values found for \bar{m} and \hat{p} we find from the graph

$$\text{Eff.}(\bar{p}) = .68.$$

It follows that we employed an inefficient method of estimation and the rejection of the above hypothesis may be groundless. We, therefore, proceed to estimate the parameters of the negative binomial law with the maximum-likelihood method. For computational purposes it is convenient to write equation (21) as

$$\frac{1}{n} \sum_{i=1}^n \lambda(r_i, \hat{p}) - 2.30258 \log_{10} \left(1 + \frac{\hat{m}}{\hat{p}} \right) = 0. \quad (21a)$$

\hat{m} is identical with $\bar{m} = .66981$ and $n = 318$. We must solve equation (21a) for \hat{p} . We find three trial values for \hat{p} in the neighborhood of $\bar{p} = .52197$, say $\hat{p} = .3, .4$, and $.5$.

The writer finds it convenient to write columns 1 and 2 of Table 3 on a strip of paper, seeing to it that the spacing of the figures is the same as the spacing in Table 1. By placing the strip next to the columns headed $\hat{p} = .3, .4$, and $.5$ of Table 1 is easy to find after continuous multiplication of the observed frequencies with the corresponding tabular values and division by 318,

$$\frac{1}{318} \sum \lambda(r, .3) = 1.25782,$$

$$\frac{1}{318} \sum \lambda(r, .4) = .98108, \text{ and}$$

$$\frac{1}{318} \sum \lambda(r, .5) = .81169;$$

and finally by substitution of appropriate values into (21a).

\hat{p}	$y = f(\hat{m}, \hat{p})$	δ	δ^2
$\hat{p}_{-1} = .3$	$y_{-1} = +.08450$	—	—
$\hat{p}_0 = .4$	$y_0 = -.00269$	—	—
$\hat{p}_{+1} = .5$	$y_{+1} = -.03330$	—	—

By inverse interpolation we may now find the root of equation (21a), i.e.

$$\hat{p} = \hat{p}_0 + uh, \quad (22)$$

where h is the unit tabulated (in this case $h = 0.1$) and

$$u = \frac{y_{-1} - y_{+1}}{2\delta^2} \pm \sqrt{\left(\frac{y_{-1} - y_{+1}}{2\delta^2}\right)^2 - \frac{2y_0}{\delta^2}}. \quad (23)$$

From the above values we have

$$\hat{p} = .40000 - .04303 \times .1 = .39570.$$

The expected frequencies for these M.L. estimates of the law (eq. 1) are given in the 6th column of Table 3.

Out of five cells used for computing total χ^2 , four show an improvement in comparison to the moment fit. Total χ^2 has been reduced to one half of its former value and, most important, $P = .18$ so that we may confidently accept the hypothesis of absence prone-ness.

Entering Figure 1 with the efficient estimates \hat{m} and \hat{p} we find from the graph

$$\text{Eff.}(\bar{p}) = .63,$$

which is not vitally different from the result based on the moment estimates. This indicates that the decision whether to proceed to a maximum-likelihood fit, i.e.

$$\text{Eff.}(\bar{p}) \geq .80,$$

is not invalidated by using the moment estimates.

(b) *Accident Proneness:*

It is a well-known fact that certain accident distributions may successfully be represented by a negative binomial law. It may be shown (Chambers and Yule, (8)) that parameter p is independent of the time exposure.

Table 4 gives the distribution of *minor* accidents of two groups of workers of the same division as mentioned in the example on absence proneness. The first group is composed of workers who had already had their full annual leave, whereas the second is made up of all those who had taken no leave. The observational period was six months. As the workers were all exposed to the same environmental risk, and as there is little likelihood that the leave group is, on the average, more prone to accidents than the non-leave group, we should expect to find similar estimates for parameter p from the two samples.

TABLE 4*

Comparison of the Method of Moments and the Method of Maximum Likelihood in Fitting the Negative Binomial in Studying Accident Proneness

Method of Estimation	Moments				Maximum Likelihood	
Number of minor accidents	With annual leave		Without annual leave		With A.L.	Without A.L.
	f_o	f_E	f_o	f_E	f'_E	f''_E
0	77	81.7	73	68.7	77.9	71.6
1	36	33.6	28	34.5	35.8	32.8
2	24	18.6	22	19.4	20.0	18.3
3	13	11.2	8	11.3	11.9	10.8
4	4	7.0	9	6.7	7.2	6.6
5	3	4.4	5	4.0	4.5	4.1
6	2	2.9	2	2.4	2.8	2.5
7	1	1.9	2	1.5	1.8	1.6
8	2	1.2	0	.9	1.1	1.0
9	2	.8	1	.6	.7	.6
10	0	.5	1	.3	.5	.4
15	1	1.2	-	.7	.8	.7
n	165	165.0	151	151.0	165.0	151.0
m	1.34545		1.33775		1.34545	1.33775
p	.59345		.80479		.69870	.69659
χ^2	3.864		3.631		2.453	3.077
P	.28		.30		.49	.39

*The brackets in table indicate grouping for χ^2 -test.

The moment estimates are

with annual leave : $\bar{p} = .593$,

without annual leave : $\bar{p} = .805$.

The two estimates of p do not seem to agree. However, entering Figure 1 with the moment estimates, we find

with annual leave : $\text{Eff.}(\bar{p}) = .61$,

without annual leave : $\text{Eff.}(\bar{p}) = .68$.

It is clear that the moment method is inefficient. The corresponding maximum-likelihood estimates are

with annual leave : $\hat{p} = .699$,

without annual leave : $\hat{p} = .697$,

showing excellent agreement. The corresponding χ^2 -tests are given in Table 4. In both cases the distributions fitted by the maximum-likelihood method represent the observations much better than do the moment graduations.

(c) *Two-Hand Co-ordination Test:*

The Two-Hand Co-ordination Test is used primarily as a measure of speed of performance. A study of the errors *per se* incurred on this test revealed differences in liability to making errors. For individuals this liability is remarkably constant from trial to trial. The split-half reliability of the errors compares favourably with the best of current performance tests. In view of the above the distribution of test errors might be expected to follow a negative binomial law. Table 5 gives the distribution of total errors for ten trials on the Two-Hand Co-ordination Test for 504 subjects.

The moment fit ($P = .10$), although acceptable, is not too satisfactory. From Figure 1 we see that for

$$\bar{m} = 7.28 \quad \text{and} \quad \bar{p} = .63$$

the efficiency of the estimation process is very low, i.e.

$$\text{Eff.}(\bar{p}) \approx 0.4.$$

We proceed to estimate the parameters by the maximum-likelihood method. The resulting χ^2 -test gives $P = .53$. Once again the

"errors of estimation" have clouded the conclusions based on the χ^2 -test. As the maximum-likelihood fit is entirely satisfactory, there exists good reason to use the negative binomial law as a mathematical model of Two-Hand Co-ordination errors.

TABLE 5*

Comparison of the Method of Moments and the Method of Maximum Likelihood in Fitting the Negative Binomial to Errors on the Two-Hand Co-ordination Test

Number of Errors	f_0 (Obs.)	f_E (Mom.)	f'_E (Max. Lik.)	Number of Errors	f_0 (Obs.)	f_E (Mom.)	f'_E (Max. Lik.)
0	74	102.1	82.1	23	1	3.3	3.3
1	58	59.3	57.5	24	3	3.0	3.0
2	51	44.5	46.1	25	1	2.7	2.6
3	49	35.9	38.6	26	2	2.5	2.4
4	42	30.0	32.9	27	2	2.3	2.1
5	23	25.6	28.4	28	3	2.0	1.9
6	30	22.1	24.7	29	3	1.9	1.7
7	20	19.3	21.6	30	3	1.7	1.6
8	17	16.9	19.0	31	1	1.5	1.4
9	18	14.9	16.7	32	0	1.4	1.2
10	16	13.2	14.8	33	2	7.5	5.2
11	11	11.8	13.1	38	2		
12	7	10.5	11.6	40	3		
13	10	9.4	10.3	42	2	7.2	5.6
14	7	8.4	9.2	44	1		
15	6	7.5	8.2	47	1		
16	5	6.8	7.3	49	1		
17	7	6.1	6.5	51	1		
18	2	5.5	5.8	68	1		
19	4	5.0	5.2	73	1		
20	5	4.5	4.6	n	504	504.0	504.0
				m		7.28373	7.28373
21	3	4.0	4.1	p		.63126	.77510
				χ^2		28.495	18.888
22	5	3.7	3.7	P		.10	.53

*The brackets in table indicate grouping for χ^2 -test.

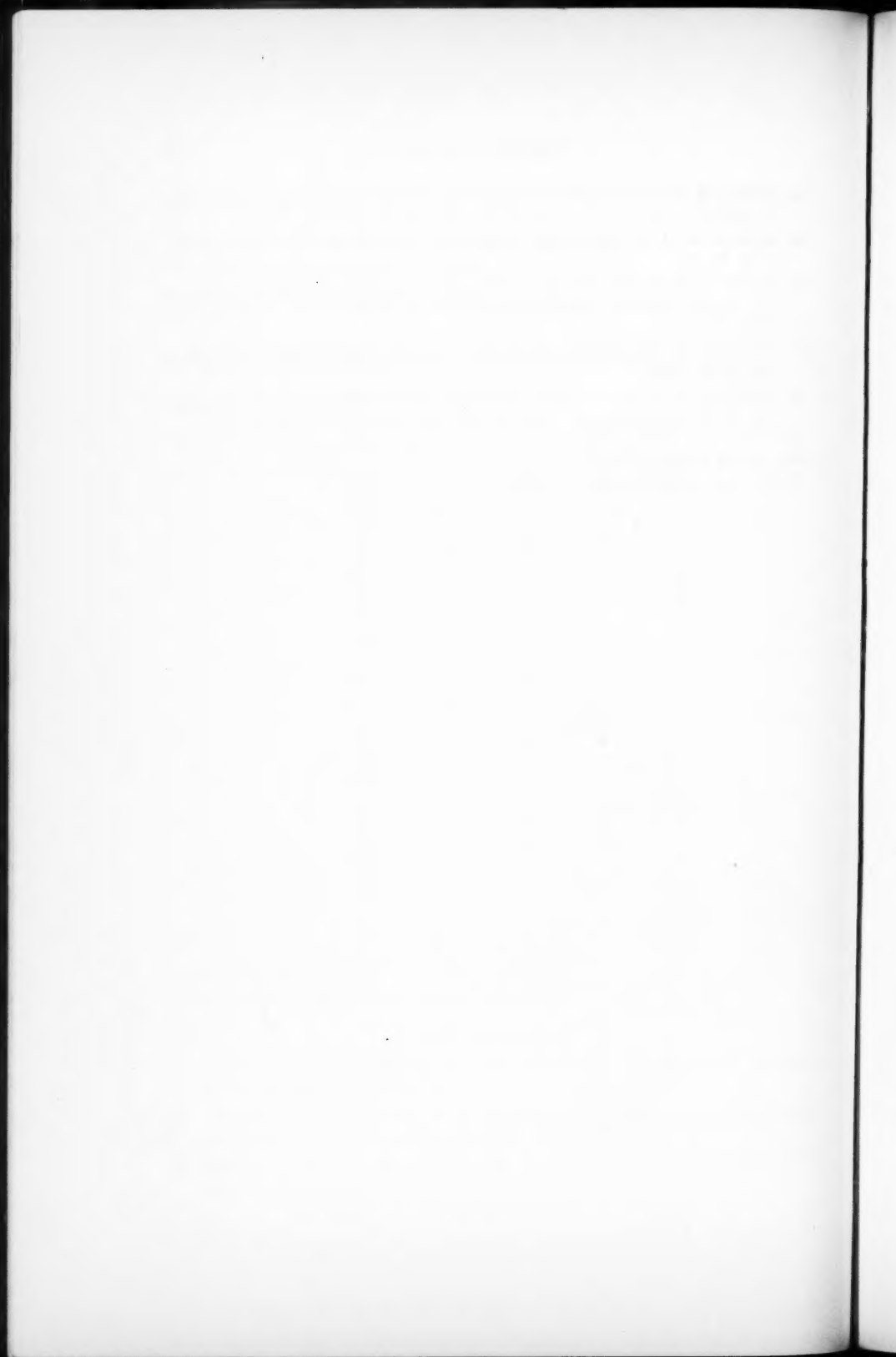
REFERENCES

1. Fisher, R. A. On the mathematical foundations of theoretical statistics. *Philos. Trans.*, 1921, A222, 309.
2. Greenwood, M., and Yule, G. U. An inquiry into the nature of frequency-distributions of multiple happenings, etc. *J. roy. stat. Soc.*, 1920, 83, 255.
3. Kendall, M. G. The advanced theory of statistics. London: Charles Griffin and Co., Ltd., 1945.

4. Fisher, R. A. The negative binomial distribution. *Ann. Eug., Lond.*, 1941, 11, 182.
5. Haldane, J. B. S. The fitting of binomial distributions. *Ann. Eug., Lond.*, 1941, 11, 179.
6. Maritz, J. S. On the validity of inferences drawn from the fitting of Poisson and negative binomial distributions to observed accident data. *Psychol. Bull.*, 1950, 47, 434.
7. Fisher, R. A. Statistical methods for research workers. Edinburgh: Oliver and Boyd, 1948.
8. Chambers, E. G., and Yule, G. U. Theory and observations in the investigation of accident causation. *Supp. J. roy. stat. Soc.*, 1941, 7, No. 2, 89.

Manuscript received 2/3/50

Revised manuscript received 7/12/50



A SUCCESSIVE APPROXIMATION METHOD OF MAXIMIZING TEST VALIDITY

GOLDINE C. GLESER AND PHILIP H. DUBOIS
WASHINGTON UNIVERSITY

The ratio of item validity to item-total correlation can be used to select items which will tend to yield the maximum correlation with a criterion. Items to be retained are identified by comparing the ratio for each item with the validity of the original test. Further improvement of the validity in the experimental sample can be obtained by adding items to or removing items from the selected nucleus, according to recomputed ratios involving the correlations of the items with the nucleus and evaluated by means of a revised cut-off point. With slight variations, the method may be used for interest and personality tests as well as for aptitude material. The principal advantage over previous methods is that for any cycle of the analysis an exact cut-off point is provided.

The problem of choosing from among a number of items that group yielding the maximum correlation with a criterion has long troubled testmakers. Since out of n items, one may use any possible combination of 2, 3, 4, ... n items, there would be $2^n - (n + 1)$ validity coefficients to compare in order to find the optimal solution empirically. Such a task is obviously prohibitive. The rational approach would be to find the multiple correlation between the items and the criterion, under the restriction that all weights be either 1 or 0. No feasible routine for this procedure has been proposed.

Various attempts have been made to find a practical solution. Horst (4) developed the "Method of Successive Residuals," which can be used with either differential or unit weights and which builds up the test by starting with the item of highest validity and then determining which of the $n - 1$ items will correlate highest with the unpredicted part of the criterion, a process which is continued until the "optimum" test is selected. The Toops L-Method (8) is another "build-up" technique for selecting either items of a test or tests for a battery using unit gross-score weights. Wherry and Gaylord (9) modified this procedure so that it can be used with differential weights as well. In all of these methods, the amount of labor is considerable.

Other methods have been devised in which some type of index

is used to determine the relative contribution of each item to the total test validity. On this basis a predetermined number of items are eliminated. Horst (5) published such a method which involves the computation of the mean criterion score and the mean total test score of all subjects who answered correctly on any particular item.

The index he obtains can be shown to be equivalent to $\frac{r_{jc}\sigma_j\sigma_c}{r_{jt}\sigma_j\sigma_t}$ where

the subscript j refers to the item, c to the criterion and t to the total test, and the correlations are point-biserials. A scatter diagram is plotted with the numerator as ordinate and the denominator as abscissa. All items are retained for which r_{jc} is positive while r_{jt} is negative, plus all items for which both correlations are positive and which lie above an arbitrary radius drawn from the origin with slope such that the predetermined number of items are excluded. This is equivalent to retaining the items with the largest index. The procedure may be repeated using the selected items as a test nucleus if further refinement is desired. Horst claims that this method is less time-consuming than his previous technique and at the same time yields higher validities for the selected group of items. Gulliksen (3) recently published a similar graphic item selection procedure in which $r_{jc}\sigma_j$ is plotted against $r_{jt}\sigma_j$ and those items retained which lie above a radius drawn from the origin. In this method no consideration is given to items having a negative correlation with the total score. It is also assumed that $\sum_j r_{jt}\sigma_j$ is nearly equal to $\sum_j r_{js}\sigma_j$ where r_{js} is the correlation between the item and the score on the selected subset of items. However the elimination of a group of items may have a considerable effect on the item-total correlation of a retained item.

Richardson and Adkins (6) have proposed an index for determining the contribution of the items based on the relative weight with which the item would enter into the prediction of the criterion when added to a test exclusive of that item. The index,

$$\frac{r_{jc} - r_{jt}r_{tc}}{(r_{tc} - r_{jc}r_{jt})\sigma_j},$$

is the ratio of the beta weight of the item to the beta weight of the test divided by the standard deviation of the item. The items having the lowest indices are dropped. In this method again no adjustment is suggested for the change in weight resulting from successive elimination of items.

Flanagan (2) has suggested a short method of item selection in which a nucleus of the most valid items is first selected, as judged by any convenient type of correlation with the criterion. Items are added to or subtracted from this nucleus by comparing the item-nucleus correlation of each item with the item-criterion correlation. Those items which have a higher correlation with the criterion score than with the nucleus are retained while those having a lower correlation with the criterion than with the nucleus are dropped. This process may be repeated until no further improvement in validity results. Flanagan remarks that the first approximation secures a very large proportion of the possible improvement. Thorndike (7) has suggested an adaptation of Flanagan's technique utilizing the partial correlation of each item with the criterion, holding constant the variability associated with the nucleus.

In determining the best technique to be used to maximize the validity of a test we should note that in every method the item indices are subject to sampling fluctuations and all error variance is weighted in favor of the test-maker. Thus there is no reason to suspect that any one of the methods, including the method developed here, is superior from the standpoint of greater stability in the resulting validity coefficient for subsequent samples. The choice of method, therefore, seems to depend mainly on the work involved for the obtained increase in validity.

A method of item selection is developed and presented here which, while it has points in common with practically all of the short methods mentioned above, has certain unique features and results in a reasonably exact solution of the problem of selecting items yielding maximum validity. This method of "successive approximations" takes into account the changes in item-total correlations which result after the first selection is made and as a consequence adds and subtracts additional items to the original selection, each additional cycle resulting in a closer approximation to the "perfect" solution. An exact cut-off point is provided for the number of items to be included at each cycle.

Consider an experimental test of n items which has been administered to N individuals. The items may be scored in any systematic manner, such as most preferable (1), neutral (0), and least preferable (-1), or simply as right or wrong. Product moment correlations are obtained between each item of the test and the criterion, between each item and the total score, and between the total score and the criterion.

Using the following notation,

c_i = the deviation of an individual's criterion score from the mean,

x_{ij} = deviation of an individual's score on a particular item j from the mean of that item, and

t_i = deviation of an individual's total score from the mean score,

we have

$$\begin{aligned}
 r_{ct} &= \frac{\sum_{i=1}^N c_i (x_{i1} + x_{i2} + \dots + x_{in})}{N \sigma_c \sigma_t} \\
 &= \frac{r_{1c} \sigma_1 \sigma_c + r_{2c} \sigma_2 \sigma_c + \dots + r_{jc} \sigma_j \sigma_c + \dots + r_{nc} \sigma_n \sigma_c}{\sigma_c \sigma_t} \\
 &= \frac{\sum_{j=1}^n r_{jc} \sigma_j}{\sigma_t}
 \end{aligned} \tag{1}$$

but

$$\sigma_t^2 = \frac{\sum_{i=1}^N (x_{i1} + x_{i2} + \dots + x_{in}) t_i}{N} = \sum_{j=1}^n r_{jt} \sigma_j \sigma_t. \tag{2a}$$

Dividing (2a) by σ_t , we find

$$\sigma_t = \sum_j r_{jt} \sigma_j. \tag{2b}$$

Substituting (2b) in (1) yields

$$r_{ct} = \frac{\sum r_{jc} \sigma_j}{\sum r_{jt} \sigma_j}. \tag{3}$$

From equation (3) it can be seen that the validity of a test is increased if the validity of any item is increased or if the correlation of any item with the total score is decreased. Accordingly, if one considers a test, all of whose items have the same difficulty and validity and the same item-total correlation, the validity of the test as a whole would equal the validity of any item divided by its correlation with the total test, and the larger this ratio, the greater the validity. It would appear, therefore, that for a group of relatively homogeneous items, those items for which this ratio is the largest would yield a test approaching the maximum validity. Items, how-

ever, differ in their means, standard deviations, and correlations with other items. By a more precise development of what happens when a single item j is removed from a test, it may be possible to estimate which items should be left in and which should be removed. Since

$$r_{c(t-j)} = \frac{\sum c_t - \sum c_j}{N\sigma_c\sigma_{t-j}} = \frac{r_{ct}\sigma_t - r_{jc}\sigma_j}{\sigma_{t-j}}, \quad (4)$$

for $r_{c(t-j)}$ to be larger than r_{ct}

$$\frac{r_{ct}\sigma_t - r_{jc}\sigma_j}{\sigma_{t-j}} > r_{ct}.$$

Multiplying both sides of the inequality by σ_{t-j} and collecting terms, we have

$$r_{ct}\sigma_t - r_{jc}\sigma_j > r_{ct}\sigma_{t-j},$$

$$r_{ct}(\sigma_t - \sigma_{t-j}) > r_{jc}\sigma_j.$$

If $(\sigma_t - \sigma_{t-j})$ is positive,

$$\frac{r_{jc}\sigma_j}{\sigma_t - \sigma_{t-j}} < r_{ct}; \quad (5)$$

while if $(\sigma_t - \sigma_{t-j})$ is negative,

$$\frac{r_{jc}\sigma_j}{\sigma_t - \sigma_{t-j}} > r_{ct}. \quad (6)$$

In order to put the quantity $(\sigma_t - \sigma_{t-j})$ into more usable form, we note that

$$\sigma_t^2 - \sigma_{t-j}^2 = \sigma_t^2 - (\sigma_t^2 - 2r_{jt}\sigma_j\sigma_t + \sigma_j^2) = 2r_{jt}\sigma_j\sigma_t - \sigma_j^2. \quad (7)$$

Since $\sigma_t^2 - \sigma_{t-j}^2 = (\sigma_t - \sigma_{t-j})(\sigma_t + \sigma_{t-j})$, it follows that

$$\sigma_t - \sigma_{t-j} = \frac{2r_{jt}\sigma_j\sigma_t - \sigma_j^2}{\sigma_t + \sigma_{t-j}} = \left(r_{jt} \frac{2\sigma_t}{\sigma_t + \sigma_{t-j}} - \frac{\sigma_j}{\sigma_t + \sigma_{t-j}} \right) \sigma_j. \quad (8)$$

$\sigma_t + \sigma_{t-j}$ very closely approximates $2\sigma_t$, so

$$\sigma_t - \sigma_{t-j} \approx (r_{jt} - \sigma_j/2\sigma_t)\sigma_j. \quad (9)$$

Substituting (9) in (5) and (6) yields

$$\frac{r_{jc}}{r_{jt} - \sigma_j/2\sigma_t} < r_{ct}, \quad \text{if } r_{jc} - \sigma_j/2\sigma_t \text{ is positive;} \quad (10a)$$

$$\frac{r_{jc}}{r_{jt} - \sigma_j/2\sigma_t} > r_{ct}, \text{ if } r_{jc} - \sigma_j/2\sigma_t \text{ is negative.} \quad (10b)$$

Thus we see that if the validity of the test is to be increased by dropping the item being considered, $\frac{r_{jc}}{r_{jt} - \sigma_j/2\sigma_t}$ for that item should be less than the validity r_{ct} if the denominator is positive, and greater than this validity if the denominator is negative.

For tests where the items are numerous and are scored as right or wrong and where the standard deviation of the total test is greater than 5, $\sigma_j/2\sigma_t$ is .05 or less and can be ignored for the first approximation of the new test form. Then the inequalities (10a) and (10b) reduce simply to

$$r_{jc}/r_{jt} < r_{ct} \quad (r_{jt} > 0) \quad (11a)$$

$$r_{jc}/r_{jt} > r_{ct} \quad (r_{jt} < 0). \quad (11b)$$

Thus if the correlation of an item with the total test is positive and the item index r_{jc}/r_{jt} is less than the validity of the test, the validity will be raised by dropping that item from the test; whereas if the correlation of the item with the total test is negative, the index should be greater than the validity of the test if the item is to be removed.

Actually we are interested in removing *several* items from a test at the same time. The above considerations would hold only when such changes have no effect on the remaining items. However, when several items are removed, the correlations of each item with the score on the nucleus of remaining items is changed somewhat. In practice, therefore, we would probably find that the removal of *all* items which do not meet the criteria indicated above will result in some items being removed which would now add to the validity of the selected items, while some which have been retained would no longer meet the recalculated criteria. For this reason the first selection should be only a trial selection, and it is necessary to recalculate all of the item sub-total score correlations based on this new nucleus of items selected by the index, and compare the new indices with a recalculated test validity.

At this point it becomes necessary to determine which items to add to a given nucleus of items in order to increase further the validity of this nucleus taken as a test. Proceeding as above and with the same notation, let us determine when an item *added* to a test will increase the validity.

$$r_{c(i+j)} = \frac{r_{ct}\sigma_t + r_{jc}\sigma_j}{\sigma_{t+j}} > r_{ct}. \quad (12)$$

$$r_{jc}\sigma_j > r_{ct}(\sigma_{t+j} - \sigma_t).$$

$$\frac{r_{jc}\sigma_j}{\sigma_{t+j} - \sigma_t} > r_{ct} \quad \text{if } (\sigma_{t+j} - \sigma_t) > 0. \quad (13)$$

By the same method used for equations (7) and (8), we find

$$\sigma_{t+j} - \sigma_j = \left(r_{jt} \frac{2\sigma_t}{\sigma_t + \sigma_{t+j}} + \frac{\sigma_j}{\sigma_t + \sigma_{t+j}} \right) \sigma_j,$$

or

$$\sigma_{t+j} - \sigma_j \approx (r_{jt} + \sigma_j/2\sigma_t) \sigma_j, \quad (14)$$

$$\frac{r_{jc}}{r_{jt} + \sigma_j/2\sigma_t} > r_{ct} \quad \text{if } r_{jt} + \frac{\sigma_j}{2\sigma_t} > 0, \quad (15a)$$

and

$$\frac{r_{jc}}{r_{jt} + \sigma_j/2\sigma_t} < r_{ct} \quad \text{if } r_{jt} + \frac{\sigma_j}{2\sigma_t} < 0. \quad (15b)$$

Thus we see that an item not included in the first selection may now be selected if its correlation with the nucleus of selected items is such that it conforms to the inequalities (15a) or (15b), while items in the nucleus may now be discarded if they conform to inequalities (10a) or (10b) where r_{ct} in this case refers to the correlation of the *selected* items with the criterion.

In light of the considerations stated above, there are two practical methods for selecting items for a test. The method to be used depends on the degree of precision desired and the type of test material to which it is applied.

Method I:

This method is suitable for tests consisting of items which are fairly homogeneous as to subject matter for which one desires to choose quickly those items which will increase the validity for a particular criterion. The items must be of the type scored right or wrong. In this case use of the index alone will increase the validity substantially. The procedure is to:

1) Obtain all of the item-criterion and item-total point-biserial correlations for each item.

2) Calculate for each item its index r_{jc}/r_{jt} .

3) For items with positive item-total correlations select those items for which the index is greater than the validity of the original form of the test.

4) If negative item-total correlations occur, then such items should be retained if the ratio r_{jc}/r_{jt} is less than the validity of the original test.

Method II:

This method is applicable to a more heterogeneous test or one for which it is desired to make a more exact selection. The procedure follows:

1) Obtain a nucleus of items as in Method I or, considering only the item validities, pick out those items which have positive and substantial correlations with the criterion. A working rule might be to use items with validities significantly greater than zero. Since further selection will be made and criteria exist for either adding or removing items, it is possible to make the first selection by this alternate method and thus save the time involved in finding all of the original item-total score correlations.

2) Score the papers on the basis of the items selected in Step (1) and obtain the correlations (r_{js} 's) of all the original items with the subtotal score and also the sub-test-criterion correlation, r_{sc} .

3) Obtain the ratio $\frac{r_{jc}}{r_{js} \pm \sigma_j/2\sigma_s}$ for each item; where the minus sign is used for those items included in the scoring (i.e., those selected in Step 1) and the plus sign is used for items not in the first selection.

4) The revised test will now consist of those items having a positive denominator for which this ratio is larger than the validity r_{sc} and those items for which the denominator of the index is negative and the index is less than r_{sc} . Some items in the first nucleus will probably be rejected, while other items not included in the first nucleus will now be accepted.

5) Steps 2, 3, and 4 may be repeated after each new selection until no further changes occur. Usually a second or third adjustment is sufficient for practical purposes. Particularly where the correlations are obtained using some type of grouping of the scores, a change in one or two items will not affect the correlations, so that the second adjustment will give the desired result within the degree of accuracy of the correlations.

This method was used on the revised Object-Aperture Test (1) consisting originally of 64 items. The test as a whole had a validity of .53 with grades in General Engineering Drawing. Selection of items on the basis of Method I resulted in a choice of 39 items with a validity of .67. On the basis of a second and third adjustment as indicated above, three items were added and two removed, resulting in the selection of 40 items with a correlation of .69 with the criterion.

In dealing with questionnaires, personality inventories, and biographical data blanks where the scoring for each item is determined empirically, the item validities will indicate both which items will enter into the initial nucleus and the direction of scoring of each item. Once the keying has been determined so that $r_{jc} \geq 0$ for all items, selection proceeds in the same fashion as indicated above.

In order to illustrate how Method II works, the following hypothetical problem is presented here. Table 1 presents hypothetical scores for 20 people on ten items and the criterion score, the latter obtained by the following formula:

$$c = \text{item 1} + 2 \text{ times item 3} + \text{item 5} + \text{item 7} + \text{item 9} + 2.$$

The correlation between the "test" and "criterion" scores is .60. The second, third, and fourth columns of Table 2 show the item-criterion correlations and the item total-score correlations and also the ratio r_{jc}/r_{jt} for each item. It can be seen that the only items for which the ratio is above .60 are items 1, 3, 5, 7, and 9. These, of course, are the items which will give a maximum correlation with the "criterion." Recomputing the validity r_{cs} for these five items gives .97. However, under the assumption that nothing was known as to the true source of the criterion, the new item-total correlations were

calculated and also the ratios $\frac{r_{jc}}{r_{jt} \pm \sigma_j/2\sigma_s}$ for each item, as shown in

the last three columns. The new indices indicate that no further changes should be made.

TABLE 1
Hypothetical Test Scores on Each Item and Criterion Scores

Subject	Items										Total Score	Criterion Score
	1	2	3	4	5	6	7	8	9	10		
S ₁	1	1	1	1	1	1	1	1	1	0	9	8
S ₂	1	1	1	0	1	0	1	1	1	1	8	8
S ₃	1	0	1	0	1	0	1	0	1	0	5	8
S ₄	0	1	1	1	1	0	1	1	1	0	7	7
S ₅	1	1	1	1	1	1	0	1	1	1	9	7
S ₆	1	1	1	0	1	0	1	1	0	1	7	7
S ₇	0	1	1	0	1	1	0	0	1	0	5	6
S ₈	1	1	0	1	1	1	1	1	1	1	9	6
S ₉	0	0	1	0	1	0	1	0	0	0	3	6
S ₁₀	1	1	0	0	1	1	1	0	0	0	5	5
S ₁₁	1	1	1	1	0	1	0	1	0	1	7	5
S ₁₂	0	1	0	1	1	1	1	0	1	0	6	5
S ₁₃	1	0	0	1	0	0	1	0	1	0	4	5
S ₁₄	1	1	0	1	1	1	0	1	0	1	7	4
S ₁₅	1	1	0	0	0	1	0	1	1	0	5	4
S ₁₆	0	0	1	1	0	0	0	0	0	0	2	4
S ₁₇	0	1	0	1	0	1	1	1	0	0	5	3
S ₁₈	1	0	0	0	0	0	0	0	0	0	1	3
S ₁₉	1	1	0	1	0	1	0	1	0	0	5	3
S ₂₀	0	1	0	1	0	1	0	1	0	0	4	2
σ _j	.48	.43	.50	.49	.49	.49	.50	.49	.50	.46		
<hr/>												
N = 20				ΣC = 106			ΣTC = 646					
ΣT = 113				ΣC ² = 626			r _{tc} = .60					
ΣT ² = 735				σ _c = 1.79								
σ _t = 2.20												

TABLE 2
Item Correlations and Ratios for Selection of Items in Hypothetical Test

Item	r_{cj}	r_{jt}	r_{jc}/r_{jt} Cut-off: .60	r_{js}	$r_{js} \pm \sigma_j/2\sigma_s$	r_{jc}
						$r_{js} \pm \sigma_j/2\sigma_s$ Cut-off: .97
1	.23	.36	.64	.33	.17	1.35
2	.03	.69	.04	.08	.23	.13
3	.73	.26	2.81	.54	.37	1.97
4	-.26	.29	-.90	-.25	-.08	3.25*
5	.76	.56	1.36	.79	.62	1.23
6	-.32	.38	-.84	-.25	-.08	4.00*
7	.54	.26	2.08	.63	.46	1.17
8	.02	.66	.08	.03	.20	.10
9	.61	.48	1.27	.68	.51	1.20
10	.32	.64	.50	.31	.47	.68

*Since Inequality (10b) applies, these items are rejected.
Using Index 1, items 1, 3, 5, 7, 9 are selected. With the test rescored on these 5 items $\sigma_s = 1.47$, $r_{cs} = .97$.

Index 2 indicates no further selection is necessary.

REFERENCES

1. DuBois, P. H., and Gleser, G. The Object-Aperture Test; a measure involving visualization in three dimensions. *Amer. Psychol.*, 1948, 3, 363 (abstract).
2. Flanagan, J. C. A short method for selecting the best combination of test items for a particular purpose. *Psychol. Bull.*, 1936, 33, 603-4.
3. Gulliksen, H. Item selection to maximize test validity. *Proceedings of the 1948 Invitational Conference on Testing Problems*, 13-17.
4. Horst, P. Item analysis by the method of successive residuals. *J. exp. Educ.*, 1934, 2, 254-263.
5. Horst, P. Item selection by means of a maximizing function. *Psychometrika*, 1936, 1, 229-244.
6. Richardson, M. W., and Adkins, Dorothy C. A rapid method of selecting test items. *J. educ. Psychol.*, 1938, 29, 547-552.
7. Thorndike, R. L. *Personnel Selection*. New York: John Wiley and Sons, Inc., 1949.
8. Toops, Herbert A. The L-Method. *Psychometrika*, 1941, 6, 249-266.
9. Wherry, R. J., and Gaylord, R. H. Test selection with integral gross score weights. *Psychometrika*, 1946, 11, 173-183.

c
t
v

l
s
l
f
l
c
s
s
P
t
i
n
e
a
g
n
e
p
w
a

BOOK REVIEW

JOHN VON NEUMANN and OSKAR MORGENSTERN. *Theory of Games and Economic Behavior*. (2nd Revised Edition) Princeton University Press, 1947, pp. xviii + 641.

A review of this book at this late date requires explanation. The reviewer is aware of the existence of a good many reviews, and he does not hesitate to admit that some of these reviews are good reviews. Good in that they distill much of the spirit of this 650-page opus in a mere 15-20 page paper of an expository nature (1), (2). As is understandable, most of the reviews were published in journals of economics or mathematics. However, the impact of this book transcends these fields: The problems it raises for scientists of any quantitative denomination are analogous to those raised for staff officers in the critique of maneuvers in which a new weapon has been tested.

With this in mind let us confess without further ado that this is a difficult book. What makes it difficult is not what is expected from the reader in terms of mathematical background or familiarity with the facts and problems of economics. The book is difficult because it expresses new ideas and uses new unfamiliar techniques to buttress these ideas. The authors try to develop a general theory of economic behavior de novo. In their concern for rigor they proceed in a manner that makes it hard for them to give the reader, through examples, a feeling for the power and the beauty of the new approach.

The authors—John von Neumann, a mathematician's mathematician and Oskar Morgenstern, a well known economist of the Austrian school—start out by stating their credo: Traditional mathematical economics has been unsuccessful because of the tools it has used. They were the tools of the differential calculus forged in the birth pangs of Newtonian Physics. Now the complexity of social phenomena is at least equal to the complexity of those encountered in Physics. "It is therefore to be expected—or feared—that mathematical discoveries of a stature comparable to that of calculus will be needed in order to produce decisive success in this field." The immediate task of social science (economics being the prototype of social science) is thus twofold: (1) continuation in the direction of the descriptive approach ("our knowledge of the relevant facts of economics is incomparably smaller than that commanded in physics at the time when the mathematization of that subject was achieved") and (2) development of a mathematical precision tool for a limited field. The authors' scholarly modesty as well as their scientific and social philosophy are expressed in these sentences: "The great progress in every science came when, in the study of problems that were modest as compared with ultimate aims, methods were developed which could be extended further and further. . . . The sound procedure is to obtain first utmost precision and mastery in a limited field, and then to proceed to another, somewhat wider one, and so on. This would do away with the unhealthy practice of applying so-called theories to economic or social reform where they are in no way

useful." This is an important point for the methodology and strategy of a good many incipient scientific disciplines.

The author's advice is then briefly this: turn away from the "burning" questions, concern with them merely delays progress. Find out as much as you can about the behavior of the individual and the simplest forms of exchange. Develop gradually a theory based on a "careful analysis of the ordinary everyday interpretation of economic facts." This is a heuristic procedure: you are just groping your way from unmathematical plausibility considerations to a formal structure. Too bad! But this is the way to proceed if you want your final theory to be mathematically rigorous and conceptually general. In its first applications your results will appear trivial since they were never in doubt. But continue, work with more complicated problems until finally you will score your real successes when you can mathematically predict what will happen.

The reviewer, who is not an economist, must confess here to a certain bewilderment. He has no doubt that such a strategy has proved successful in the natural sciences where controlled experimentation led theory down a primrose path. He feels, however, less sanguine about the possibility of keeping "burning" or controversial issues out of the construction-job marked "theory of economics." The economic facts of life seem much too interwoven with the behavior—rational or otherwise—of the individual and of society. The simple recipe of chanting "an economic fact is a fact" may not constitute a powerful enough incantation to dispel the next fellow's or the next society's plausibility considerations. The authors were undoubtedly aiming at scientific neutrality. Still at least one reviewer "doubts whether their method based essentially on a capitalist form of production covers all rational economics" (3). It is safe to say that the goals and needs of a society interact with the building of an economic theory in a complex manner.* The way in which the Theory of Games has caught on in the fields of economic and military strategy makes it relatively safe to state that this mathematical structure too is keeping its appointment with burning questions (4). These somewhat critical remarks do not detract from the intrinsic value of the book and from the real enjoyment that is felt by the serious reader willing to dig through a prose heavily loaded with footnotes and references to earlier sections.† The student will soon find himself fascinated by the ease with which combinatorics, set theory and linear algebra are developed under the very nose of unsuspecting penny pitchers and poker players.

The Theory of Games approaches economic theory from the viewpoint of the individual. It must therefore make certain assumptions concerning his motives. The authors do not hesitate to accept the traditional view according to

*Some of these interaction problems were considered in recent papers read before the Boston meetings of the Institute for the Unity of Science. See in particular the papers by Dr. A. Kaplan on "Scientific Method and Social Policy" and Prof. Philipp Frank on "The Logical and the Sociological Aspects of Science." Dr. Kaplan was mainly concerned with the role of perspective, programmatic and methodic scientism while Prof. Frank attempted to analyze the extra-scientific factors responsible for the acceptance of a particular theory.

†A sample from section 15.4.3 (page 119): "The interpretation which we are now going to give to the result of 13.5.3 is based on our considerations of 14.2-14.5—particularly those of 14.5.1., 14.5.2—and for this reason we could not propose it in 13.5.3.

which the consumer wants to obtain a maximum of satisfaction and the entrepreneur a maximum of profits. Once maximization of utility has been stated as the principle of rational behavior a further assumption is necessary before we can manipulate the variable "utility" numerically (for simplicity's sake we might for instance decide to use monetary units to measure utility). We must "accept the picture of an individual whose system of preferences is all-embracing and complete, i.e., who for any two imagined events (or combination of events with stated probabilities) possesses a clear intuition of preference." In their axiomatic treatment of utility von Neumann and Morgenstern combine *this* condition of a complete system of preferences with the condition of transitivity of preference relations into the single concept of complete ordering.* Our authors emphasize that they are dealing only with utilities experienced by one person with no implications concerning the comparison of utilities belonging to different individuals. Nobody should therefore expect to simply open the book in order to find weighting functions that would permit him to determine the utility function for a social group. In this connection von Neumann and Morgenstern state that the social maxim of "the greatest possible good for the greatest possible number" is self-contradictory, since "a guiding principle cannot be formulated by the requirement of maximizing two or more functions at once."

We are now almost ready to take a look at what constitutes a solution, i.e., a set of rules for a participant in an economic game. Since we cannot explore all possible types of games let us first see if we cannot categorize the types of possible economic situations an individual might encounter. A brief summary of these categories resembles a primitive system of counting: one, two, many. In the Robinson Crusoe economy the mathematics is, theoretically, simple: there are a certain number of wants and a certain number of commodities and the problem is to obtain maximum satisfaction. Obviously an ordinary (though admittedly multi-variable) maximum problem.—Now take the case of two participants in a social exchange economy. This case turns out to have particular significance in the formulation of the whole theory. It has still certain elements in common with a maximum problem but certain radically new features have been added. Each participant attempts to maximize a function of which he does not control all the variables.

As the number of players on the economic stage increases beyond two a new concept comes to the fore, the concept of *coalition*. By means of coalitions we can attempt to reduce a more complex exchange economy to what is essentially a two-person game, but this task is neither easy nor can it always be carried out convincingly. The hope still persists that, as in physics, it will some day be easier to deal statistically with an economy of 150 million people than with the problems involving exchanges between the butcher, the baker, and the candlestick maker. But the authors strongly insist that "only after the theory for moderate number of participants has been satisfactorily developed will it be possible to decide whether extremely great numbers of participants simplify the situation." They stress

*The axiomatic treatment leaves utility a number determined up to a linear transformation. For a discussion of the relation of transformation groups to psychological scales see the Chapter by S. S. Stevens on "Mathematics, Measurement and Psychophysics" in the forthcoming *Handbook of Experimental Psychology* (S. S. Stevens, editor).

that the analogy with the celestial vs. statistical mechanics situation tends to be fallacious. There the *general* theory of the mechanics of several bodies is well known. The difficulties that stem from the special, computational application of the theory to, let us say, the solar system are greater than those encountered in predicting the overall behavior of for instance 1025 freely moving particles.

We have now seen under what circumstances the concept of utility can be handled numerically and we have some ideas concerning the types of economic situations we have to investigate. The aim of the inquiry is to find the mathematically complete principles that define "rational behavior" for the participants in a social economy. While these principles ought to be perfectly general, it might be easier to start out by finding solutions for certain characteristic special cases. The next question is: how will we recognize a solution when we see one? Here is the intuitively plausible concept of a solution according to our authors: Each participant must have a set of rules telling him how to behave in *every* situation that may arise (in other words these rules make allowance for irrational behavior on the part of others).

This is a point at which the great similarity between economics and the "everyday concept of games" is driven home. Games become now formally mathematical models for social and economic problems. They constitute ideal theoretical constructs: they are amenable to precise, exhaustive, and not too complicated definitions; they further bear a resemblance to reality in the traits judged *essential* for the purposes at hand. The solution of the game which is derived from these constructs is in general an involved combinatorial catalogue. Its summary for the individual answers the question of *how much* he will get if he behaves "rationally." This is the minimum he can get; if others make mistakes, he gets more.

In particularly simple games the solution will consist of a single imputation, i.e., a single statement as to how the total proceeds are to be distributed among the participants. However, as soon as we get into more complicated games our solution undergoes parthenogenesis: the single imputation is replaced by a set of imputations: It turns out that this set of imputations is not ordered; in other words no single imputation is superior to ("dominates") all others. To our authors this lack of transitivity is a most typical phenomenon in social organizations. If the dominance relations between various states of society are of a cyclical nature (B is superior to A , C is superior to B and finally A is superior to C ; compare this with the "paper form" of horses, or baseball teams if you wish), then we have not only different possible equilibrium positions but also a possibility of passing from one of these equilibrium states to another.

This brings us to the next important point which is the static nature of this theory. The authors are aware of the fact that a dynamic theory would be more complete and therefore preferable. Yet they feel that it is futile to try to build such a dynamic theory as long as the analysis of equilibrium states is not yet thoroughly understood. This static character of the theory is of course a very serious handicap if these models are to be used for the study of adaptive or learning processes.

Much of what we have discussed up to this point comes from the introductory chapter. Chapter 2 furnishes a general formal, set-theoretical description of games of strategy. Many of the important terms are defined in these 40 pages. The unifying concept of a player's strategy emerges: it constitutes the plan of ac-

tion which specifies what choices he will make in every possible situation.

This formal model is now put to work. If we leave Robinson Crusoe playing solitaire on his island, the simplest of the remaining games is the zero-sum two-person game. Zero-sum games are an important classification of all possible games. As the name implies the sum of all payments involved, by all players at the end of the game, is always zero. Two-person games are simple because of the absence of coalitions. Under these circumstances the main problems can be formulated as follow: How does a player plan his strategy? How much information does he possess and what role does the amount of information play in determining his moves? The zero-sum two-person game constitutes in other words a good dry run for the entire theory.

Let us see what happens in this two-person game. Jim's moves are determined by the rules of the game and by his desire to win as much as possible. But he proceeds cautiously; he assumes that his strategy has already been found out by his opponent Joe. This is obviously the worst that could have happened to Jim. Jim chooses therefore a strategy that will assure him a gain that is not less than a certain amount (or a loss not greater than a certain amount). If Joe actually is not as smart as Jim gives him credit for, Jim will be better off than he anticipates.

The core of the zero-sum two-person game is constituted by the Min-Max (or Minimax) problem. With its help the authors show that games in which perfect information prevails (like chess, for example) are particularly rational or strictly determined. For these games permanently optimal strategies exist. In this sense if the theory of chess were really fully known there would be nothing left to play.

Then what about non-strictly determined games like Matching Pennies? Is there any hope that the "common-sense" behavior of players will yield a clue to a solution? Since it is hard to find out the intentions of your opponent, the next best thing in this game is to concentrate on avoiding having your own intentions found out. This can be done by playing a statistical or "mixed strategy" (play "tails" or "heads" with a probability mixture of 50:50) to protect yourself against loss. Our solution can therefore be couched in terms of mixed strategies.

Before we go on to further theoretical considerations we get a chance to play a few elementary games like Stone, Paper, Scissors. We also watch Sherlock Holmes escape Professor Moriarty by helping him pick a good strategy. And as a special test we get initiated into the intricacies of a rather formalized stud poker. The emphasis is on bluffing.

With the theory of the zero-sum two-person game as our base of operations we attack the zero-sum three-person game. The analysis is here dominated by the concept of coalitions: genesis, internal arrangements and understandings, strength and stability. From here we go on to the general treatment of the zero-sum n -person game, and finally to the most general type of game by removing the restriction of the zero-sum. Here we leave the realm of games played for entertainment to enter the realm of economic reality since the sum of all payments, or the social product, is in general different from zero.

A non-zero-sum game of n persons can be shown to be reducible to a zero-sum $(n+1)$ -person game by the introduction of a fictitious player, Mr. $(n+1)$. He turns out to be a handy mathematical gadget and a sad character at the same

time. He obligingly permits us to make the sum of the amounts obtained by the players equal to zero by picking up the total check. In order to be able to do this he must have no influence whatsoever on the course of the game and remain excluded from all transactions connected with the game.

We have now run the whole economic gamut from Robinson Crusoe's simple maximum problem to Mr. $(n + 1)$'s strange market place. In the remaining chapter von Neumann and Morgenstern discuss generalizations of their concepts of solution, domination and utility.

The reviewer does not want to imply that all readers will get to these ultimate extensions of the theory. He feels that the reader will feel well rewarded if he works through the first five chapters, i.e., up to and including the zero-sum three-person game. Most of the fundamental ideas are developed in these 240 pages, including the classic Minimax problem and the treatment of coalitions. He may by that time be able to think of some relevant applications of the theory which has unfolded before him or he may have gained enough courage to strike out in new directions.

The reviewer does not feel competent to evaluate the promise the book holds for the future, nor to prescribe it as a wonder drug to those who are dealing with difficult quantitative problems in this area. He feels, however, that the effect of the von Neumann-Morgenstern opus will not be that of strong medicine but rather that of a beneficial catalyst in the thinking of social scientists.

REFERENCES

- 1 HURWICZ, L., *Amer. econ. Rev.*, 1945, 35, 909-925.
- 2 MARSHAK, J., *J. pol. Econ.*, 1946, 54, 97-115.
- 3 GUMBEL, E. J., *Ann. Amer. Acad. pol. soc. Sci.*, 1945, 239, 209-210.
- 4 McDONALD, J., A theory of strategy, *Fortune*, 1949, June. pp. 100-110.

Psycho-Acoustic Laboratory
Harvard University

Walter A. Rosenblith

- LAWRENCE E. ABT AND LEOPOLD BELLAK (editors). *Projective Psychology*. New York: Alfred A. Knopf, Inc., 1950, pp. xvii + 485 + xiv.
- C. S. BLUEMEL. *War, Politics, and Insanity*. Denver: The World Press, Inc., 1948, pp. 121.
- NANDOR FODOR, AND FRANK GAYNOR, (editors) *Freud: Dictionary of Psychoanalysis*. New York: Philosophical Library, 1950, pp. 208+xii.
- M. C. GRECO, *Group Life: The Nature and Treatment of Its Specific Conflicts*. New York: Philosophical Library, 1950, pp. 357+xvi.
- J. P. GUILFORD, *Advanced Statistical Exercises*. Beverly Hills: Sheridan Supply Co., 1950, pp. 110+ii.
- HAROLD GULLIKSEN. *Theory of Mental Tests*. New York: John Wiley & Sons, 1950, pp. 486+xix.
- TJALLING C. KOOPMANS (ed.) *Statistical Inference in Dynamic Economic Models*. New York: John Wiley & Sons Inc., 1950, pp. 438+xiv.
- J. NEYMAN, *First Course in Probability and Statistics*, New York: Henry Holt & Co., 1950, pp. 350+ix.
- ARTURO ROSENBLUETH. *The Transmission of Nerve Impulses at Neuroeffector Junctions and Peripheral Synapses*. Published jointly by the Technology Press of Massachusetts Institute of Technology and John Wiley & Sons, New York, 1950, pp. xiv+325.
- SAMUEL A. STOUFFER et al. *Measurement and Prediction*, Princeton: Princeton University Press, 1950, pp. 756 + x.
- JOHN D. TRIMMER, *Response of Physical Systems*, New York: John Wiley & Sons, 1950, pp. 268+vii.
- ABRAHAM WALD, *Statistical Decision Functions*. New York: John Wiley & Sons, Inc., 1950, pp. 179+ix.
- MARY H. WEISLOGEL. *Procedures for Evaluating Research Personnel with a Performance Record of Critical Incidents*. Pittsburgh: American Institute for Research, 1950, pp. 42+iv.

THE HISTORY OF THE

THE HISTORY OF THE

THE HISTORY OF THE

THE HISTORY OF THE

THE HISTORY OF THE

THE HISTORY OF THE

THE HISTORY OF THE

THE HISTORY OF THE

THE HISTORY OF THE

THE HISTORY OF THE

THE HISTORY OF THE

THE HISTORY OF THE

THE HISTORY OF THE

PSYCHOMETRIC MONOGRAPHS

A new issue of Psychometric Monograph Series, No. 5, *The Description of Aptitude and Achievement Tests in Terms of Rotated Factors*, by John W. French, will be ready for distribution later this year. The issue will contain about 300 pages and the price will be \$4.00.

Previously published issues of the Psychometric Monograph Series are as follows:

Thurstone, L. L. Primary mental abilities. *Psychometric Monograph No. 1*, \$2.00.

Thurstone, L. L. and Thurstone, Thelma Gwinn. Factorial studies of intelligence. *Psychometric Monograph No. 2*, \$1.50.

Wolfe, Dael. Factor analysis to 1940. *Psychometric Monograph No. 3*, \$1.25.

Thurstone, L. L. A factorial study of perception. *Psychometric Monograph No. 4*, \$2.50.

Orders for any issue should be sent to:

THE UNIVERSITY OF CHICAGO PRESS
5750 Ellis Avenue
Chicago, Illinois

The Psychometric Monograph Committee is composed of J. P. Guilford, *Chairman*; L. L. Thurstone, Harold Gulliksen, Paul Horst, and Frederic Kuder. Manuscripts and correspondence for this series should be addressed to:

J. P. GUILFORD, *Chairman*
Psychometric Monograph Committee
Box 1134
Beverly Hills, California



